Homogenization and trend estimation of GNSS water vapour observations and atmospheric reanalyses. Application to climate change analysis.

Khanh-Ninh Par

Université Paris Cité

Institut de Physique du Globe de Paris

École doctorale Sciences de la Terre et de l'environnement et physique de l'Univers
STEP'UP n°560
Équipe de géodésie

Homogénéisation et estimation des tendances des observations GNSS de la vapeur d'eau et des réanalyses atmosphériques. Application à l'analyse du changement climatique.

Homogenization and trend estimation of GNSS water vapour observations and atmospheric reanalyses. Application to climate change analysis.

Par Khanh-Ninh Nguyen

Thèse de doctorat de Sciences de la Terre

Dirigée par Olivier Bock
et Emilie Lebarbier

Présentée et soutenue publiquement le 8 Décembre 2023

Devant un jury composé de :

Mme Hélène Brogniez, *PU, LATMOS, Université de Versailles Saint-Quentin-en-Yvelines, rapporteur*
M. Athanassios Argiriou, *Full professor, University of Patras, rapporteur*

M. Hervé Douville, *Directeur de Recherche, Météo-France, Université Paul Sabatier, Examinateur*
M. Clément Narteau, *Professeur, IPGP, Université Paris Cité, Examinateur*

Mme Emilie Lebarbier, *PU, Université Paris Nanterre, co-directrice de thèse*
M. Olivier Bock, *Directeur de Recherche, IGN, IPGP, Université Paris Cité, directeur de thèse*

Acknowlegdement

First and foremost, I extend my deepest gratitude to my supervisors, Olivier Bock, Emilie Lebarbier, and Françoise Guichard, for their invaluable guidance, patience, and unwavering support throughout my PhD journey, from preparation to completion. Special appreciation goes to Olivier, a professional researcher, for his consistent and continued support and tolerance. His countless feedback and meticulous guidance have been pivotal in refining my academic writing and sharpening my arguments.

I am deeply thankful to Emilie Lebarbier for her rigorous mathematical perspective and her wealth of creative ideas. Our weekly meetings kept me on track and progressively shaped my academic growth. Francoise Guichard, though no longer with us, provided warm support at the outset of my PhD journey during the challenging times of COVID quarantine. Her guidance in both scientific and French cultural aspects has left an indelible mark on my development.

I also could not have undertaken this journey without my PhD defense committee members: Prof. Athanassios Argiriou, Prof. Hélène Brogniez, Dr. Hervé Douville, and Prof. Clément Narteau. I am grateful for your acceptance of the invitation to be part of my committee, and for your insightful feedback and constructive critiques, which have significantly enhanced this thesis.

My appreciation goes out to the Institut de Physique du Globe de Paris (IPGP) and Institut national de l'information géographique et forestière (IGN) for their financial support and for providing an environment conducive to research and innovation. This opportunity has been pivotal in my academic and personal development.

Additionally, I owe a great deal to my wonderful lab mates, whose support has been a constant source of motivation. A heartfelt thanks to Maylis, Ana, and Marie for welcoming me warmly to the team, sharing your experiences, and providing encouragement and support both academically and personally. My gratitude also extends to Louis-Marie, Charlotte, Laetitia, Boris, Jiao for their companionship throughout this journey, and to Paul, Kevin, Julien, Kristel, Laurent, and Gwendoline for their invaluable support and advice in research and academic matters. Thanks also to Anne and Nora for their assistance with administrative tasks.

Lastly, but most importantly, a heartfelt thank you to my family. Your unwavering love, understanding, and encouragement have been the pillars of my strength. To my mom, dad, Duc, and Quan: Thank you for everything. This PhD thesis is dedicated to you.

Résumé *(en français)*

La vapeur d'eau joue un rôle important dans le bilan énergétique et le cycle hydrologique de la Terre. Avec le réchauffement climatique, la vapeur d'eau dans l'atmosphère a tendance à augmenter, mais la répartition géographique des tendances de la vapeur d'eau dans les modèles climatiques et dans les observations reste assez incertaine. Les estimations de la vapeur d'eau intégrée (IWV) fournies par le système mondial de navigation par satellite (GNSS) constituent une nouvelle source de données qui peut servir à confirmer les tendances observées et simulées de la vapeur d'eau. Des études précédentes ont montré l'existence d'inhomogénéités dans les enregistrements IWV du GNSS dues aux changements des instruments de mesures des stations et dans les méthodes de traitement des données, et ont recommandé leur homogénéisation pour les applications climatiques.

Comme les approches récentes d'homogénéisation de données climatiques, cette étude utilise une méthode statistique de segmentation pour détecter les changements abrupts appelées ruptures dans la moyenne des séries de différences IWV entre les données GNSS et les données de réanalyse. Les séries de différences sont utilisées pour atténuer la forte variabilité temporelle inhérente aux séries IWV, rendant difficile la détection de petits sauts sur les séries brutes. Cette étude utilise le package R GNSSseg développé par Quarello (2020).

Le premier objectif de cette thèse était d'étudier la sensibilité des résultats de la segmentation et des estimations de tendances aux propriétés des données d'entrée. Deux jeux de données GNSS spécifiques ont été considérés (IGS repro1 et CODE REPRO2015) ainsi que deux réanalyses différentes (ERA-Interim et ERA5). Il s'avère que seuls 45 à 50% de ruptures communes sont détectées entre les deux jeux de données GNSS. La longueur de la série ou l'utilisation de données auxiliaires dans le traitement des données GNSS IWV a un impact plus faible, avec 70 – 80% de détections similaires. Les résultats de la segmentation sont sensibles au niveau de bruit et à la présence d'un biais périodique dans les données, qui sont principalement dus à des différences de représentativité entre les observations GNSS et les réanalyses. Les données récentes de GNSS et de réanalyses présentent des différences moins importantes et permettent la détection de sauts plus faibles. Les estimations des tendances résultantes sont sensibles au nombre et à la position des ruptures. C'est pourquoi la validation des ruptures détectées est une étape cruciale avant l'estimation des tendances. Il a été constaté que les changements d'instruments GNSS disponibles sous forme de métadonnées peuvent expliquer 35% des ruptures, en laissant 65% non documentés ou correspondants à des inhomogénéités dans les réanalyses.

Le deuxième objectif de cette thèse était de développer une méthode de classification automatique de post-détection pour distinguer les ruptures du GNSS de la réanalyse. La méthode "d'attribution" proposée combine les données GNSS et de réanalyse d'une station principale avec des données similaires de stations voisines. Chaque paire "station principale-station voisine" comprend quatre séries de base à partir desquelles six séries de différences sont formées. Ensuite, pour chaque rupture détectée dans la série principale, un test de significativité du saut de moyenne à cet instant est effectué, et une règle statistique de prédiction est construite pour attribuer la rupture testée au GNSS ou à la réanalyse. Lorsque plusieurs stations proches sont disponibles, une solution combinée est proposée. L'aspect original de la méthode d'attribution développée est l'utilisation d'une méthode d'inférence généralisée des moindres carrés, pour tenir compte de l'hétéroscédasticité et de l'autocorrélation dans les données, et d'une méthode d'apprentissage automatique. La principale nouveauté réside dans la manière dont la règle de prédiction est construite, en utilisant les résultats des tests obtenus à partir des données réelles

avec une procédure de rééchantillonnage. Une analyse de sensibilité montre que les prédictions dépendent d'une manière ou d'une autre de la stratégie de rééchantillonnage, en utilisant des échantillons équilibrés ou déséquilibrés, du niveau de significativité du test et également du niveau de bruit dans les données. Un facteur limitant dans les réseaux éparses est le bruit élevé quand les séries sont très éloignées. Lorsque la méthode est appliquée aux données GNSS CODE REPRO2015 et à la réanalyse ERA5, 62% des ruptures testées sont attribuées au GNSS, 19% à la réanalyse et 10% à des changements à la fois dans le GNSS et la réanalyse.

Mots-clés :vapeur d'eau intégrée, GNSS, homogénéisation, tendance, segmentation, attribution, réanalyse, régression linéaire généralisée, classification supervisée

Abstract *(in english)*

Water vapour plays an important role in the Earth's energy balance and hydrological cycle. As the climate warms, water vapour in the atmosphere tends to increase, but the geographical distribution of water vapour trends in climate model simulations and observations remains quite uncertain. Global Navigation Satellite System (GNSS) estimates of Integrated Water Vapor (IWV) are a new data source that may serve for the verification of observed and simulated water vapour trends. Previous studies have reported inhomogeneities in the GNSS IWV records due to changes in the station instrumentation and in data processing methods, and recommended their homogenization for climate applications.

Following modern approaches to climate data homogenization, this study uses a statistical segmentation method to detect abrupt changes, called change-points, in the mean (or jumps) of IWV differences series between GNSS and reanalysis data. Difference series are used to mitigate the strong temporal variability inherent in the IWV series which makes the detection of small jumps difficult on the raw series. This study employs the GNSSseg R package developed by Quarello (2020).

The first objective of this thesis was to investigate the sensitivity of the segmentation results and trend estimates to the input data properties. Two specific GNSS datasets were considered as inputs (IGS repro1 and CODE REPRO2015) and two different reanalysis products (ERA-Interim and ERA5). It is found that only 45-50% similar change-points are detected when the input data are altered. Altering the series length or the auxiliary data used in the processing of GNSS IWV data has a smaller impact, with 70-80% similar detections. Segmentation results are sensitive to the noise magnitude and the presence of a periodic bias in the input data, mainly due to representativeness differences between the GNSS observations and reanalysis. More recent GNSS and reanalysis products have smaller differences and allow the detection of smaller jumps. The subsequent trend estimates are sensitive to the number and position of change-points. Therefore, validation of detected change-points is a crucial step before trend estimation. It is found that changes in GNSS instrumentation available as metadata can explain 35% of change-points, leaving 65% as either undocumented GNSS changes or changes in the reanalysis data.

The second objective of this thesis was to develop an automatic classification method, operated as a post-processing step, to distinguish GNSS and reanalysis change-points. The proposed "attribution" method combines GNSS and reanalysis data from a main station with similar data from nearby stations. Each main-nearby pair comprises four base series from which six series of differences are formed. Then, for each detected change-point in the main series, a significance test is applied to the mean before and after the change-point, and a classifier or a statistical predictive rule is constructed to attribute the tested change-point to GNSS or reanalysis. When several nearby stations are available, a combined solution is proposed. The original aspect of the developed attribution method is the use of a generalized least-squares inference method, to account for heteroscedasticity and autocorrelation in the data, and a machine learning classifier. The main novelty is the way in which the predictive rule is constructed, using test results obtained from the real data with a resampling procedure. Sensitivity analysis shows that the prediction results depend somehow on the resampling strategy, using balanced or imbalanced samples, on the test significance level, and also on the magnitude of the noise in the data. A limiting factor in spare networks is the large noise with long distance between the main and nearby stations.

When applied to the CODE REPRO2015 GNSS data and ERA5 reanalysis, 62% of the tested change-points are attributed to GNSS, 19% to the reanalysis, and 10% to changes in both GNSS and reanalysis.

Keywords: integrated water vapor, GNSS, homogenization, trend, segmentation, attribution, reanalysis, generalized least square, supervised classification

Résumé substantiel *(en français)*

Introduction

La vapeur d'eau joue un rôle important dans le bilan énergétique et le cycle hydrologique de la Terre. En tant que gaz à effet de serre, elle amplifie le réchauffement de la planète par son mécanisme de rétroaction positive. La compréhension et la quantification de l'évolution de la vapeur d'eau dans le contexte du changement climatique représentent un défi scientifique majeur, comme le soulignent des recherches récentes (Colman and Soden, 2021; Douville et al., 2022). À mesure que le climat se réchauffe, la vapeur d'eau dans l'atmosphère a tendance à augmenter, mais la répartition géographique des tendances de la vapeur d'eau dans les simulations de modèles climatiques et les observations reste assez incertaine (Hersbach et al., 2020; Flato et al., 2013). Dans ce contexte, le système mondial de navigation par satellite (GNSS) offre des estimations de la vapeur d'eau intégrée (IWV) avec une haute résolution temporelle disponible dans toutes les conditions météorologiques (Guerova et al., 2016), qui peuvent servir de source indépendante de données pour la vérification des tendances de la vapeur d'eau observées et simulées (Parracho, 2017). Néanmoins, l'homogénéité des données GNSS est sensible à divers facteurs, y compris aux changements des instruments de mesures, des méthodes de traitement des données, et environnementaux (Vey et al., 2009; Bock et al., 2010; Ning et al., 2016). Dans ce contexte, cette thèse vise à contribuer à l'effort d'homogénéisation des enregistrements GNSS IWV, ce qui est essentiel pour permettre des estimations précises des tendances de la vapeur d'eau atmosphérique.

Bilan et objectifs

L'homogénéisation fait référence au processus d'identification et de correction des inhomogénéités (artificielles) dans des séries de données. Suivant les approches modernes de l'homogénéisation des données climatiques, une méthode statistique de segmentation pour détecter les changements abrupts, appelées ruptures, dans la moyenne des séries de différences IWV entre les données GNSS et les données de réanalyse (approche qualifiée de "relative") a été utilisée. L'utilisation de séries de différences est stratégique pour atténuer la forte variabilité temporelle inhérente aux séries IWV, rendant difficile la détection de petits sauts sur les séries brutes. La méthode de segmentation utilisée dans cette étude a été développée par Quarello (2020) et dédiée à l'analyse de ces séries de différences (en s'adaptant à ces caractéristiques). Elle a été validée par des études de simulations et a été reconnue comme l'une des plus efficaces lorsqu'elle a été comparée à d'autres méthodes sur des données benchmark dans l'étude de Van Malderen et al. (2020). Toutefois, la robustesse, la sensibilité et les limites de cette méthode par rapport à diverses caractéristiques des données et ses effets sur l'estimation des tendances n'avaient pas été étudiées. Il s'agit d'un aspect crucial pour son application pratique à des ensembles de données réelles. Le premier objectif de cette thèse est donc de combler cette lacune. Le second est de déterminer les origines des ruptures, qui peuvent provenir de données GNSS ou de données de réanalyse. Cette étape est appelée attribution. L'attribution précise des ruptures à leur source est essentielle avant de corriger les données IWV brutes de leurs sauts. Représentant un défi fondamental de l'approche relative, ce problème d'attribution demeure une limite clé des méthodes d'homogénéisation existantes. En réponse, cette thèse introduit une méthode automatique en post-traitement à la segmentation. Une étude poussée à la fois sur des simulations mais aussi sur les vraies données est menée afin de bien comprendre les effets des différents facteurs mis en jeu lors de cette étape d'attribution. Enfin, le dernier objectif est de corriger les séries temporelle GNSS IWV des sauts associes aux ruptures qui leurs sont attribuées.

Sensibilité de la méthode de segmentation aux propriétés des données et étude de l'impact de ces propriétés sur les estimations de tendances

Une importante étude de la sensibilité de la méthode de segmentation aux propriétés des données a été réalisée, ainsi que l'évaluation de l'impact de ces propriétés sur les estimations de tendances. Cette analyse a comparé des paires de jeux de données selon quatre facteurs critiques: les méthodes de traitement des données GNSS (IGSrepro1 vs. CODE REPRO2015), la longueur des séries (17 ans vs. 25 ans), les données auxiliaires utilisées dans la conversion de la vapeur d'eau intégrée (IWV), et les sources des données de référence (ERA-Interim et ERA5). L'accent a été mis sur des propriétés spécifiques des données, telles que les valeurs moyennes, les niveaux de bruit et les biais périodiques, car ces propriétés ont un impact sur les résultats de la segmentation, en particulier en termes d'influence sur le nombre et le positionnement des ruptures. Les résultats indiquent des impacts significatifs sur les résultats de la segmentation lorsque l'on modifie le traitement GNSS et la réanalyse de référence. Seules 45 à 49% des ruptures sont similaires dans ces cas, contre 71 à 81% pour les deux autres facteurs (la longueur de la série et les données auxiliaires). Les changements notables dans le traitement GNSS comprennent des changements dans la correction ZHD a priori, le modèle d'étalonnage de l'antenne/radome et la fonction de cartographie. Les améliorations apportées au traitement CODE permettent de réduire le bruit et le biais périodique. De même, le passage de l'ERA-Intérim à l'ERA5 en tant que référence réduit les erreurs de représentativité, ce qui entraîne là aussi une réduction du bruit et des biais périodiques, facilitant ainsi la détection des « petites » ruptures. Le taux de validation des ruptures détectées par les métadonnées est entre 30% et 35% pour tous les ensembles de données, ce qui signifie que l'impact des propriétés des données sur ce taux est faible. Ce résultat suggère 65% des ruptures détectées sont soit des changements GNSS non documentés, soit des changements dans les données de réanalyse.

L'impact sur les estimations des tendances est étudié selon deux aspects. Premièrement, l'influence des quatre facteurs précédents sur les estimations des tendances de l'IWV est étudiée. Il apparaît que la modification de la longueur de la série temporelle a l'impact le plus fort, à la fois sur la moyenne et sur la dispersion des estimations des tendances à travers le réseau. Le changement de moyenne semble indiquer l'intensification du cycle de l'eau au cours de la décennie 2010-2020 par rapport à la décennie précédente. La dispersion réduite de la tendance avec la période plus longue est principalement le résultat de l'erreur standard plus petite des estimations.Deuxièmement, l'effet de la correction des inhomogénéités sur les estimations des tendances est étudié. Cet effet est particulièrement marqué lorsque seules les ruptures validées par les métadonnées GNSS (changements de récepteur, d'antenne, de radôme) sont corrigées. Cet impact est évident dans les tendances moyennes globales, la dispersion et les différences RMS par rapport à ERA5. La correction utilisant toutes les ruptures a également un impact sur les estimations des tendances, mais dans une mesure légèrement moindre. Ceci met en exergue la nécessité d'une étape d'attribution afin de retenir uniquement les ruptures attribuables au GNSS pour la correction. Enfin, la dispersion associée aux différentes estimations homogénéisées des tendances peut être utilisée comme mesure de l'incertitude dans l'estimation des tendances à une seule station. Elle s'élève à $0.1 - 0.2$ kg m^{-2} décennie^{-1}, ou 0.5-1% décennie^{-1}, ce qui confirme la faisabilité de la détection des tendances climatiques mondiales et régionales pertinentes avec les données GNSS IWV.

Développement de la méthode d'attribution

Le second objectif de la thèse porte sur le développement d'une méthode automatique d'attribution des ruptures. Elle a pour objectif de déterminer l'origine des ruptures détectées par la segmentation, i.e. de distinguer les ruptures du GNSS de celles de la réanalyse. Cette méthode consiste à combiner les données GNSS et de réanalyse de la station principale avec des données similaires provenant d'une ou plusieurs stations voisines. Chaque paire "station principale-station voisine" comprend quatre séries de base à partir desquelles six séries de différences sont formées. On propose et développe alors la stratégie suivante: chaque rupture détectée dans la série principale, un test de significativité d'un saut de moyenne au même instant est effectué dans les autres séries de différences formées, puis une règle statistique de prédiction permet d'attribuer la rupture testée au GNSS ou à la réanalyse. Lorsque plusieurs stations proches sont disponibles, une solution combinée est proposée.

Les nouveautés de notre approche résident dans (1) la prise en compte d'une dépendance temporelle évidente et bien connue pour le test de significativité du saut à l'aide d'une méthode d'estimation GLS dans un modèle de régression, et (2) du développement d'une règle de prédiction basé sur un algorithme supervisé connu.

Pour (1), la méthode consiste à caractériser dans un premier temps l'hétéroscédasticité et l'autocorrélation présentes dans le bruit des six séries de différences. Les modèles de bruit testés sont le bruit blanc, AR(1), MA(1) et ARMA(1,1). Il est constaté que seul un petit ensemble de séries se conforme au modèle de bruit blanc. Par conséquent, prendre en compte cette dépendance, une approche GLS, connue pour fournir des estimateurs statistiquement plus efficaces, est employée afin d'évaluer la signification des sauts. Le test se fait dans un modèle de régression prenant en compte l'hétéroscédasticité et le modèle de dépendance alors identifiée du bruit mais aussi le biais périodique dans la moyenne.

Pour (2), une règle prédictive est ensuite construite en formant un classificateur pour prédire l'origine d'une rupture à partir des résultats des tests obtenus sur un ensemble de jeu de données réelles. Deux difficultés principales se sont présentées: la réponse n'est pas connue sur les données réelles et le jeu de données est de faible taille et n'est clairement pas exhaustif pour espérer construire une bonne règle de prédiction. Pour relever ces défis, nous proposons de générer un ensemble de données synthétiques basé sur les résultats des tests des données réelles en utilisant une technique de rééchantillonnage.

Lorsque la méthode est appliquée aux 81 stations GNSS du jeu de données CODE REPRO2015, avec la réanalyse ERA5 utilisée comme référence et le jeu de données NGL GNSS pour les stations voisines, 62% des ruptures testées sont attribuées au GNSS, 19% à la réanalyse et 10% à des changements à la fois dans le GNSS et la réanalyse. La majorité des points de changement identifiés est attribuée au GNSS ; cela constitue un résultat favorable et est conforme aux attentes. Les résultats de la prédiction sont sensibles à la stratégie de rééchantillonnage. Nous avons testé plusieurs variantes, en utilisant différents niveaux de signification pour les tests, des échantillons équilibrés ou déséquilibrés et différentes règles d'agrégation lorsque plusieurs stations proches sont disponibles. Les résultats des tests varient également en fonction du niveau de bruit, qui est corrélé à la distance spatiale entre la station principale et les stations voisines. Les niveaux de bruit augmentent considérablement avec la distance entre la station principale et les stations voisines dans un réseau peu dense. Ce bruit élevé peut entraver les tests de signification dans les quatre séries non colocalisées, ce qui a inévitablement

un impact sur le résultat final de la prédiction.

Conclusions et perspectives

En résumé, cette thèse contribue à la résolution de deux problèmes. Premièrement, elle a permis de mieux comprendre la sensibilité de la méthode de segmentation GNSSseg et des estimations de tendances aux propriétés des données GNSS et de référence (réanalyse). Deuxièmement, une méthode innovante d'attribution automatique a été développée, ce qui constitue une étape cruciale, avec la segmentation, au sein du processus global d'homogénéisation.

Plusieurs éléments de notre approche d'homogénéisation pourraient être améliorés et présenter des opportunités pour de futures recherches. Il s'agit notamment des éléments suivants:

1. L'approche d'attribution pourrait bénéficier d'un ensemble de données GNSS plus conséquent et plus exhaustif provenant de NGL repro3. Il serait possible d'affiner les critères de sélection des stations voisines, notamment en privilégiant des séries plus proches et avec moins de lacunes. Cela réduira sans aucune doute l'ampleur du bruit dans les séries de différences et améliorera la puissance du test du saut de moyenne. En outre, la taille plus importante de l'ensemble de données pourra permettre d'établir une nouvelle règle de classification plus performante, car si le nombre de changements testés est plus élevé, il y aura moins de répétitions dans l'ensemble de données synthétique construit par rééchantillonnage.
2. Un package R de la méthode d'attribution sera développé, permettant son accessibilité à l'ensemble de la communauté scientifique.
3. Avec le jeu de données NGL repro3, il sera aussi possible de tester d'autres approches de la méthode de segmentation mise en œuvre actuellement, comme l'utilisation de stations GNSS voisines comme référence au lieu d'une réanalyse. Lorsque la station principale et les stations voisines sont traitées de manière cohérente, les inhomogénéités restantes proviendraient principalement de changements spécifiques à la station, tels que des changements instrumentaux ou des changements dans l'environnement. Par conséquent, cette approche peut faciliter l'attribution des ruptures et conduire à des améliorations significatives dans les séries homogénéisées.

Dans le contexte du traitement des données GNSS, la segmentation s'avère très sensible aux petites variations dans la moyenne de la série. Cette sensibilité est influencée par plusieurs aspects clés du traitement des données GNSS, avec un accent particulier sur des facteurs tels que: la correction ZHD a priori, le modèle d'étalonnage de l'antenne/radome, la fonction de cartographie et l'angle de coupure de l'élévation. La cohérence et la fiabilité des produits GNSS générés par différents logiciels et incorporant diverses caractéristiques de traitement peuvent faire l'objet d'un examen minutieux.

Dans une perspective de recherche à long terme, la réalisation d'une analyse des tendances homogénéisées de la valeur de référence du GNSS offre une opportunité convaincante, en particulier dans le contexte des études sur le changement climatique. Cette analyse apporte des preuves empiriques pour répondre à une question clé: Comment la vapeur d'eau évoluera-t-elle dans un climat plus chaud ? La rétroaction de la vapeur d'eau est souvent estimée à l'aide d'un modèle climatique global (MCG), qui dépend fortement de la paramétrisation. Les jeux de données GNSS homogénéisées peuvent être comparés et utilisés comme référence pour le réglage des

paramètres dans les modèles climatiques, ce qui permet d'obtenir les "meilleures" estimations de la rétroaction de la vapeur d'eau et de ses incertitudes.

Glossary

ACF	Autocorrelation function
AR	Autoregressive model
ARMA	Autoregressive Moving Average model
CART	Classification and Regression Tree
CMIP	Coupled Model Intercomparison Project
CODE	Center for Orbit Determination in Europe
ERA-Interim	ECMWF Reanalysis - Interim
ERA5	ECMWF Reanalysis version 5
GNSS	Global Navigation Satellite System
IERS	International Earth Rotation and Reference Systems Service
IGS	International GNSS Service
IPCC	Intergovernmental Panel on Climate Change
IWV	Integrated Water Vapor
JPL/NASA	NASA's Jet Propulsion Laboratory
k-NN	k Nearest Neighbors
LDA	Linear Discriminant Analysis
MA	Moving Average model
MERRA-2	Modern-Era Retrospective analysis for Research and Applications Version 2
NGL	Nevada Geodetic Laboratory
PACF	Partial autocorrelation function
PW	Precipitable Water
RF	Random Forest
ZHD	Zenith Hydrostatic Delay
ZTD	Zenith Tropospheric Delay
ZWD	Zenith Wet Delay

Contents

Chapter 1

Introduction

1.1 Background and context

1.1.1 The role of water vapour in a changing climate

Water vapour plays a central role in Earth's energy and water cycle in a variety of ways. Being relatively transparent to the (shortwave) solar radiation and mainly opaque to the (longwave) terrestrial radiation, it is the strongest greenhouse gas in the atmosphere and contributes to nearly 75% of the total greenhouse effect (Bengtsson, 2010). It is also a key actor of energy transport, by absorbing latent heat during evaporation over the oceans and releasing it through condensation in clouds. This exchange of latent heat is pivotal in driving atmospheric circulation at various scales, from moist convection in individual thunderstorm cells to the long-range planetary circulation and moisture transport (Schneider et al., 2010; Sherwood et al., 2010). The strength of moist processes such as convection and precipitation is tightly controlled by the amount of water vapour in the atmosphere, where high amounts of moisture can lead to extreme events (e.g. heavy precipitation and floods). Another fundamental property of water vapour is it short residence time in the atmosphere of about 8 days (Trenberth, 1998), which makes it extremely variable, both spatially and temporally, and thus challenging to observe and simulate in atmospheric models.

One of the fundamental controls of water vapour in the climate system is the Clausius-Clapeyron (CC) law which relates saturation vapour pressure e_s to temperature, T (Peixoto and Oort, 1996):

$$\frac{de_s}{dT} = \frac{Le_s}{R_v T^2}. \tag{1.1}$$

where R_v is the gas constant for water vapour (461 J K^{-1} kg^{-1}), and L represents the latent heat of vapourisation (2.5×10^6 J kg^{-1}). If we integrate (1.1) between T_0 and T, we obtain:

$$e_s(T) = e_0 \exp\{-\frac{L}{R_v}(\frac{1}{T} - \frac{1}{T_0})\}. \tag{1.2}$$

where $e_0 = e_s(T_0)$.

Because temperature decreases at a mean rate of $\sim$ 6.5 K km^{-1} in the troposphere, the saturation vapour pressure varies roughly by 4 orders of magnitude between the surface and the tropopause. This steep vertical

gradient plays a major role in convection. Moisture in a raising air parcel, even if not initially saturated, can thus quickly reach the saturation level above which the latent heat released by condensation will enhance its buoyancy and may ultimately lead to deep convection. This mechanism is a major humidification process of the upper troposphere and the lower stratosphere.

The actual amount of water vapour in the atmosphere can be quantified by different variables. Relative humidity (RH) is the quantity most used for general applications (Peixoto and Oort, 1996). It is defined as $U = e/e_s$, the ratio of vapour pressure, e, of an individual air parcel at a temperature T, to saturation vapour pressure of air at the same temperature, $e_s(T)$. A collection of saturation vapour pressure formula, e_s as a function of T, can be found here (http://cires1.colorado.edu/ voemel/vp.html). Specific humidity is another widely used quantity in meteorology which expresses the fraction of mass of water vapour in an air parcel at temperature T and pressure P to its total mass: $q = \rho_v/\rho$, where $\rho_v = e/R_vT$ and $\rho = P/RT$ are the volumic masses (densities) of water vapour and the mixture of dry and air and water vapour, respectively. Relative humidity differs both qualitatively and quantitatively from other moisture variables (Peixoto and Oort, 1996). The distribution of humidity and temperature as a function of height are fundamental for many atmospheric phenomena, e.g. initiation of convection. They can be measured in-situ by balloon-borne sensors (radiosondes). Because it is constrained by the Clausius-Clapeyron law, specific humidity is quickly decreasing with height, i.e. most of the water vapour in an atmospheric column is concentrated in the lower troposphere (about 90% lies below 5 km). For this reason, the Total Column Water Vapour (TCWV), also referred to as Integrated Water Vapor (IWV) or Precipitable Water (PW), is another relevant variable often used to quantify the amount of water vapour in the atmosphere. TCWV and IWV are defined as the vertical integral of water vapour density in the atmospheric column:

$$\mathrm{IWV} = \int_0^\infty \rho_v(z)dz. \tag{1.3}$$

where $z = 0$ refers to the surface and $z = \infty$ to the top of atmosphere. IWV is expressed in units of kg m^{-2} and represents the total mass of water vapour in an unit-area column of atmosphere. A typical mi-latitude summer value would be IWV = 25 kg m^{-2}. PW is defined as the integral of $\rho_v(z)/\rho_l$, where ρ_l=1000 kg m^{-3} is the density of liquid water. It is usually expressed in mm or cm and can be interpreted as the equivalent height of liquid water if all the water vapour in an unit-area column of atmosphere would be condensed and precipitated. IWV is a quantity that can be measured by Global Navigation Satellite System (GNSS) and microwave radiometers.

Figure 1.1, borrowed from Parracho et al. (2018), shows the geographical distribution of mean IWV in boreal winter and summer from ERA-Interim reanalysis and GNSS stations. IWV reaches its maximum around the equator, ranging from about 20 to 60 kg m^{-2}, due to intense evaporation over the oceans caused by warm sea surface temperatures and strong surface winds. At higher latitudes and over the continents, IWV decreases because of limited moisture supply over land masses and lower water holding capacity of the cooler atmosphere. Another noteworthy observation is that the reanalysis is able to captures strong spatial gradients in good agreement with GNSS data.

In the context of climate change, water vapour plays a significant role in enhancing global warming through a positive feedback mechanism (Bony et al., 2006; Colman and Soden, 2021). This feedback mechanism acts in

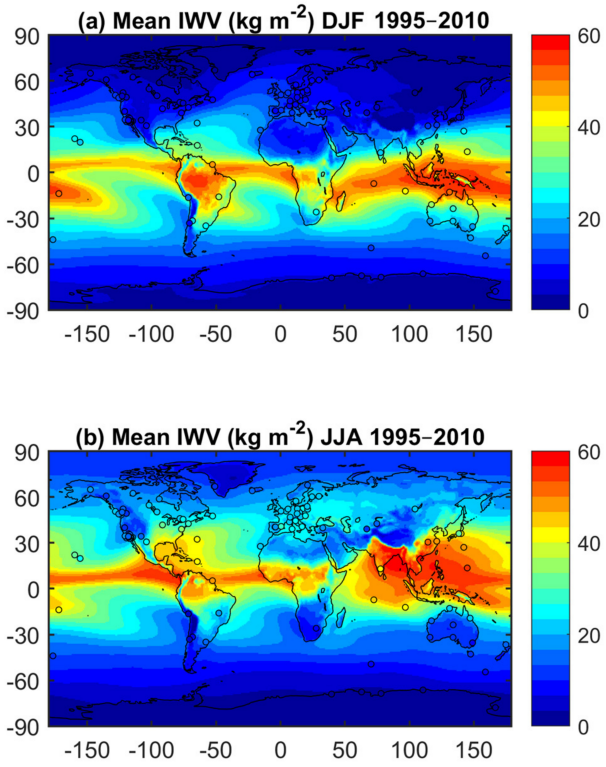


Figure 1.1 – Mean value of IWV from ERA-Interim between 1995 and 2010 for December-January-February (DJF) (a) and June-July-August (JJA) (b). Filled circles correspond to IWV retrieved by Global Positioning System (Parracho et al., 2018).

response to the forcing effect of other greenhouse gases, such as CO_2. Climate model simulations have shown that, in the scenario of doubling CO_2, temperature increase is amplified by a factor of 2 to 3 due to the water vapour feedback compared to a "no-feedback" scenario (Manabe and Wetherald, 1967; Held and Soden, 2000). The mechanism is the following: an increase in non-condensable, long-lived, greenhouse gases (e.g. CO_2, CH_4 or NO_2), induces a rise in temperature through greenhouse radiative forcing. Guided by the Clausius-Clapeyron (CC) law, this temperature increase results in a proportional elevation of saturation vapour pressure. Given that water vapour itself is a greenhouse gas, an increase in atmospheric water vapour enhances the absorption of both longwave and shortwave radiation while reducing outgoing terrestrial radiation. This, in turn, amplifies the warming effect. Moreover, the intensification of the water cycle, including heavy precipitation and flood events, is closely linked to the increase in low-altitude water vapour (Douville et al., 2021).

A key question is, what actual changes in humidity and temperature can we expect in a warming climate? (Colman and Soden, 2021) Global mean surface temperature in the first two decades of the 21st century (2001–2020) was 0.99 [0.84 to 1.10] °C higher than 1850–1900 (Gulev et al., 2021, p. 320). Guided by the CC law, the corresponding rate of increase in saturation vapour pressure at T=273 K is $\sim 7\%\ K^{-1}$, but how much is the actual water vapour changing?

Both observations and climate models have consistently supported the assumption that relative humidity remains relatively close-to-unchanged on a global scale in response to warming (Held and Soden, 2006; Trenberth et al.,

2013; Hartmann et al., 2013; Colman and Soden, 2021; Douville et al., 2022). Unchanged RH implies an increase in specific humidity and in IWV, which was also confirmed in IWV observations from various data sources (Hartmann et al., 2013; Bock et al., 2014; Wang et al., 2016; Parracho et al., 2018).

Figures 1.2 and 1.3 show the global mean (90°S-90°N) lower-tropospheric temperature (LTT) anomalies and the near global (60°S-60°N) IWV anomalies over the past decades from observations and reanalyses (Dunn et al., 2023). Both temperature and and IWV show a long-term tendency towards higher values, superposed to quite large interannual to decadal variability. The linear trend estimates of the various temperature data sets of Figure 1.2 are in the range 0.17-0.22 K decade^{-1} for the period 1979-2022 (Dunn et al., 2023, p. S36). The trends from the three reanalyses (ERA5, MERRA-2, and JRA55) are in the range 0.13-0.23 K decade^{-1} for the more recent period, 1995-2021. The corresponding IWV trends are ranging from 1.3 to 1.6% decade^{-1} and the CC scaling factors take values from 6.9 to 7.9% K^{-1} [O. Bock, personal communication]. They align closely to the 7% K^{-1}predicted by the Clausius-Clapeyron equation at T = 273 K, and thus confirm that the global mean water vapour has increased at nearly constant RH over the past 2.5 decades.

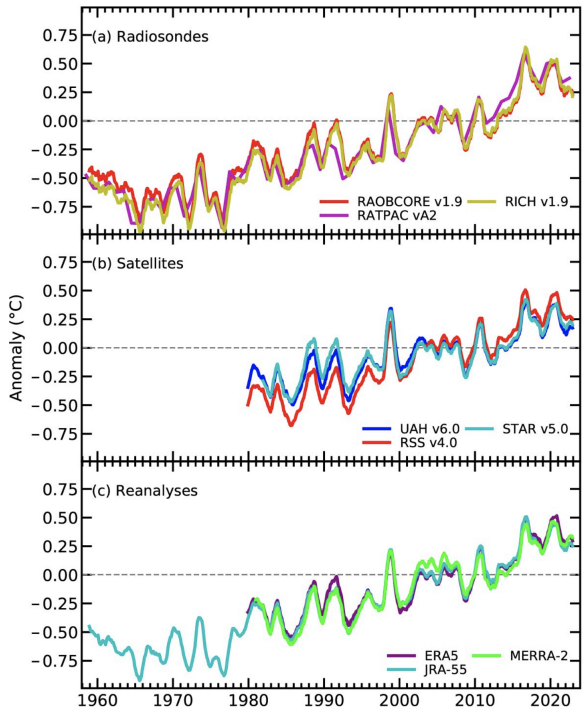


Figure 1.2 – Monthly average global lower-tropospheric temperature (LTT) anomalies (°C; 1991–2020 base period) for (a) radiosonde, (b) satellite, and (c) reanalysis datasets. Time series are smoothed using a 12-month running average. Annual averages are displayed for the Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC) dataset (Fig. A2.7 from Dunn et al., 2023).

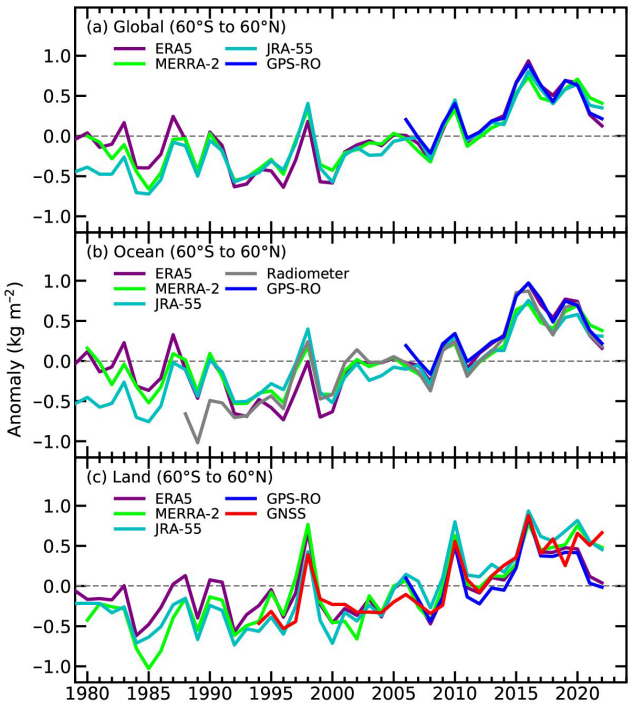


Figure 1.3 – Global mean total column water vapour annual anomalies (kg m $^{-2}$) over (a) land and ocean, (b) ocean only, and (c) land only from observations and reanalyses (ERA5, MERRA-2, JRA-55). The shorter time series from the observations have been adjusted so that there is zero mean difference relative to the ERA5 results during their respective periods of record (Fig. 2.25 from Dunn et al., 2023).

1.1.2 Consistency and homogeneity of observations, reanalyses and climate models

While a global upward trend in IWV has been deemed very likely, it is essential to evaluate the level of uncertainty in the trend estimates and understand the origin of discrepancies between the various data sources. Figure 1.3 reveals differences between the various observational techniques, as well as between the various reanalyses. The latter depart from each other especially in years before 1995 where the number of assimilated data is much lower than in more recent years. Changes in the geographical distribution and instrumentation of stations over time are a major source of spatial and temporal inhomogeneity for reanalyses (Hersbach et al., 2020) and their capacity for estimating decadal trends has been questioned (Thorne and Vose, 2010). Nevertheless, reanalysis data are extensively used to complement observational records for the evaluation of climate model simulations (Boucher et al., 2020).

In the framework of the IPCC Fifth Assessment Report (AR5), a large dispersion in IWV and lower tropospheric temperature (LTT) trends was evidenced among the Fifth Coupled Model Intercomparison Project (CMIP5) climate model simulations in tropical regions (20°S-20°N). Although all simulated trends align well on a 5.9% K^{-1} CC line (fairly consistent with CC and the assumption of constant RH), the magnitudes of their LTT and IWV trends ranged from 0.1 to 0.5 K decade^{-1} and 0.5 to 2.5% decade^{-1}, respectively (Flato et al., 2013, Fig. 9.9, p. 774). On the opposite, the trends associated with MERRA-2 and ERA-Interim reanalyses, and with two satellite products, are relatively far from the CC line, and thus rather suggest a long-term change in relative humidity. It is not known whether these discrepancies are due to remaining inhomogeneity in the observational data and/or reanalysis results, or due to problems with the climate simulations. Similar discrepancies have been reported with the more recent CMIP6 model simulations, ERA5 reanalysis, and satellite observations by Santer et al. (2021).

In a study by Allan et al. (2022), a comparison was drawn among satellite observations, ERA5 reanalysis, and CMIP6 climate model simulations. For the model simulations, they distinguished the *historical* (coupled) simulations and *amip* (atmosphere-only simulation) experiments for a subset of models. They found median IWV trend estimates of 1.9 and 1.1% decade^{-1} for the two types of simulations (period 1988-2014) compared to 0.78% decade^{-1} for ERA5 and 1.0% decade^{-1} for observations which combined infrared and microwave satellite data and ground-based humidity observations from stations. The *amip* simulations exhibited quite consistent trends with the observations. In contrast, CC scaling factors of the *historical* simulations (5.6% K^{-1}) showed better agreement with observations (5.5% K^{-1}) and ERA5 (5.8% K^{-1}) than the *amip* simulations (4.8% K^{-1}).

More recently, Douville et al. (2022) investigated the feasibility of constraining climate model projections of global mean IWV with the help of observations and reanalysis data using a Bayesian statistical method. Constraining the forced response of historical CMIP model simulations with global mean surface temperature observations since 1850 and global mean IWV data from GNSS observations and reanalyses after 1994, they could reduce the spread in global mean IWV projected to the end of the 21st century by 39%. The 5-95% confidence interval of the constrained CC scaling factors was also reduced (6.5 to 7.6% K^{-1}), with a median close to 7% K^{-1}. The projections also indicate a strong geographical disparity in the water vapour change, with a multi-model mean exceeding 12 kg m^{-2} in the tropics or 50% in desertic and polar regions, which may be

connected to deviations from CC-scaling, especially over land where constant RH cannot be maintained due to lack moisture supply from the surface despite the increasing evaporative demand. In a similar study using surface RH measurements, constrained model projections show an inevitable continental drying, especially in the northern midlatitudes (Douville and Willett, 2023).

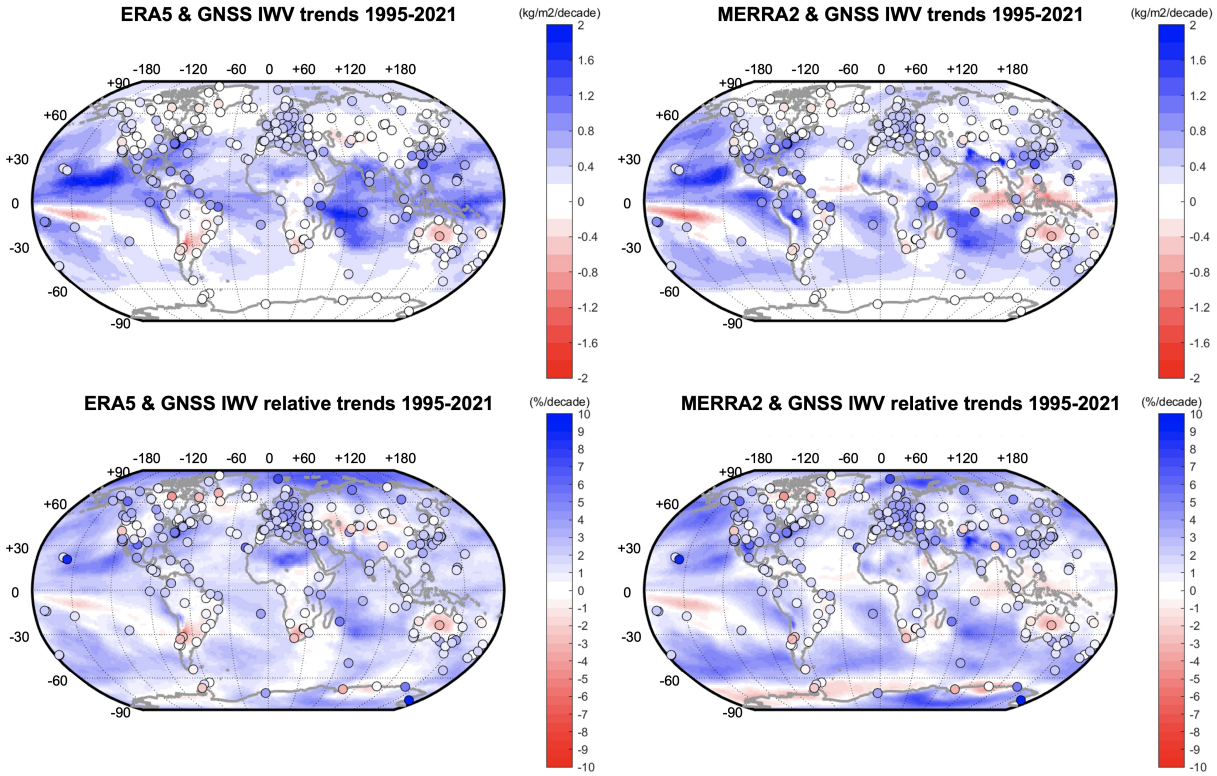


Figure 1.4 – Absolute (top) and relative (bottom) trend of the IWV. The map shows results from ERA5 (left) and MERRA-2 (right) reanalyses, while filled circles correspond to GNSS stations (courtesy O. Bock, 2023).

Figure 1.4 presents both absolute and relative trends of monthly IWV anomalies obtained at 240 GNSS stations (Bock, 2020b), alongside the MERRA-2 and ERA5 reanalyses, using the same data set as for Figure 1.3. It is important to note that, although there is a general positive trend observed across all datasets, substantial variations in the magnitude of the trends is observed between the reanalyses in various regions , with sometimes opposite signs, especially over the maritime continent, but also in the Atlantic Ocean, South America, South Asia, and Antarctica. One may note that the trend expressed in kg m^{-2} $decade^{-1}$ is very close to zero in Antarctica because the mean IWV is usually very low there. However, the relative trend (in % $decade^{-1}$) shows both positive and negative values, but with relatively poor consensus between ERA5 and MERRA-2. Parracho et al. (2018) also noted pronounced disparities between ERA-Interim and MERRA-2 over Antarctica, except at the location of the GNSS stations. Although GNSS data are not assimilated in these reanalyses, these GNSS stations are located in places where other meteorological measurements are collected and are probably assimilated.

Discrepancies between GNSS and the ERA-Interim at a number of stations were reported by Parracho et al. (2018). Inspection of the IWV time series for these stations revealed inhomogeneities in the GNSS data, in

the form of jumps and drifts, which were identified as instrumentation malfunctioning and/or replacement. Compared to Parracho et al. (2018)'s results, Figure 1.4 shows much better agreement between reanalyses and GNSS (where reanalyses agree), suggesting that the updated GNSS product used in Figure here is more homogeneous than the older one (this point will be confirmed in Chapter 3. However, inhomogeneities in the reanalyses cannot be excluded. For example, (Schroeder et al., 2016) detected jumps in IWV from reanalyses and satellite observations, and noted that several change-points in the reanalyses coincided with changes in the observing system (e.g. start and end of assimilation of satellite data in the reanalyses).

1.1.3 Inhomogeneities in GNSS IWV time series

In the context of atmospheric water vapour monitoring, GNSS offers reliable observational data with a high temporal resolution available under all weather conditions (Guerova et al., 2016). As long as these observations are not assimilated in reanalyses, they remain useful independent data for evaluating the quality of the latter, as well as climate model outputs (Parracho, 2017). Nevertheless, the homogeneity of GNSS data is susceptible to various factors, including instrumental changes, processing variations, and environmental influences (Vey et al., 2009; Bock et al., 2010; Ning et al., 2016). Jumps in IWV series of 1 to 2 kg m $^{-2}$ have been observed which are susceptible to obscure most of the underlying climatic trends which are in the range of -2 to +2 kg m^{-2} decade^{-1} at stations (Figure 1.4). Homogenization of GNSS IWV series appeared as a necessity for climate applications and the first consistent effort to this end was undertaken during the COST GNSS4SWEC project (Bock O. Pacione R. Ahmed F. Araszkiewicz A. BaŁdysz et al., 2020). Several available tools used by the climate community for the homogenization of temperature and precipitation data were tested in a benchmarking exercise organized in the framework of this project (Van Malderen et al., 2020). At the same time, Quarello (2020) developed a specialized method for the detection of change-points in GNSS minus reanalysis IWV difference series, which proved to be one of the best tested by Van Malderen et al. (2020). This thesis builds on the detection tool proposed by Quarello (2020) to include it in a full homogenization method. Before going into further details of this method, we explain the reason why GNSS minus reanalysis difference series need to be considered.

The top plot in Figure 1.5 illustrates the daily IWV GNSS series recorded at the ALIC station in Alice Springs, Australia. The IWV series at this station displays a pronounced temporal variation, with values ranging from under 5 kg m $^{-2}$ in austral winter to over 50 kg m $^{-2}$ in summer. Such substantial seasonal variability can pose challenges in detecting small jumps in the raw IWV time series. The same is observed with temperature time series and other climatic variables (Easterling and Peterson, 1995; Mitchell and Jones, 2005). To address this issue, following the "relative segmentation" approach developed by climatologists, we differentiate the GNSS IWV series with respect to a reference series. The lower plot in Figure 1.5 thus shows the IWV difference series between the GNSS and ERA5 reanalysis data extracted at the location of the station. The strong seasonal variation seen in the raw IWV series is clearly reduced, thereby improving the detectability of small inhomogeneities. The figure also presents the segmentation results (change-points are represented by vertical dashed lines) obtained with the GNSSseg method developed by Quarello (2020) (the mathematical aspects of this method will be explained in subsection 2.2.1). However, it is important to notice that, although significantly reduced, the seasonal variation in the difference series is not entirely removed. Additionally, the day-to-day scatter the difference series also exhibits strong seasonality. Both features are due to representativeness differences between the GNSS point observations and the model grid cells (Bock and Parracho, 2019), and are susceptible to induce false change-

point detections (Bock et al., 2019). In Quarello (2020)'s segmentation method, these features are included in the mathematical model in the form of a periodic bias and a monthly variance, respectively. The estimates of these two features are shown in Figure 1.5 as the purple line for the bias and the cyan line for the monthly variance.

At station ALIC, the segmentation tool detected five change-points, marked by the vertical dashed lines. They delimit changes in the mean, drawn as the broken red line, which are in the range 0.5 - 1.5 kg m^{-2}. Since the segmentation is conducted on the difference series, these inhomogeneities could originate from either the GNSS or the ERA5 data. Metadata documenting instrumental changes at the GNSS station can serve as a valuable source for validating the detected change-points. This information is reported on Figure 1.5 for the ALIC stations. Small colored triangles indicate the time position of instrumental (receiver, antenna, and/or radome) changes as well as the position of a change in the processing procedure (blue triangle). Among the five detected change-points, two coincide well with equipment changes: a receiver change in 2000 and a receiver plus antenna change in 2011. However, the other change-points are a bit far from known equipment changes to be surely attributed to GNSS origin.

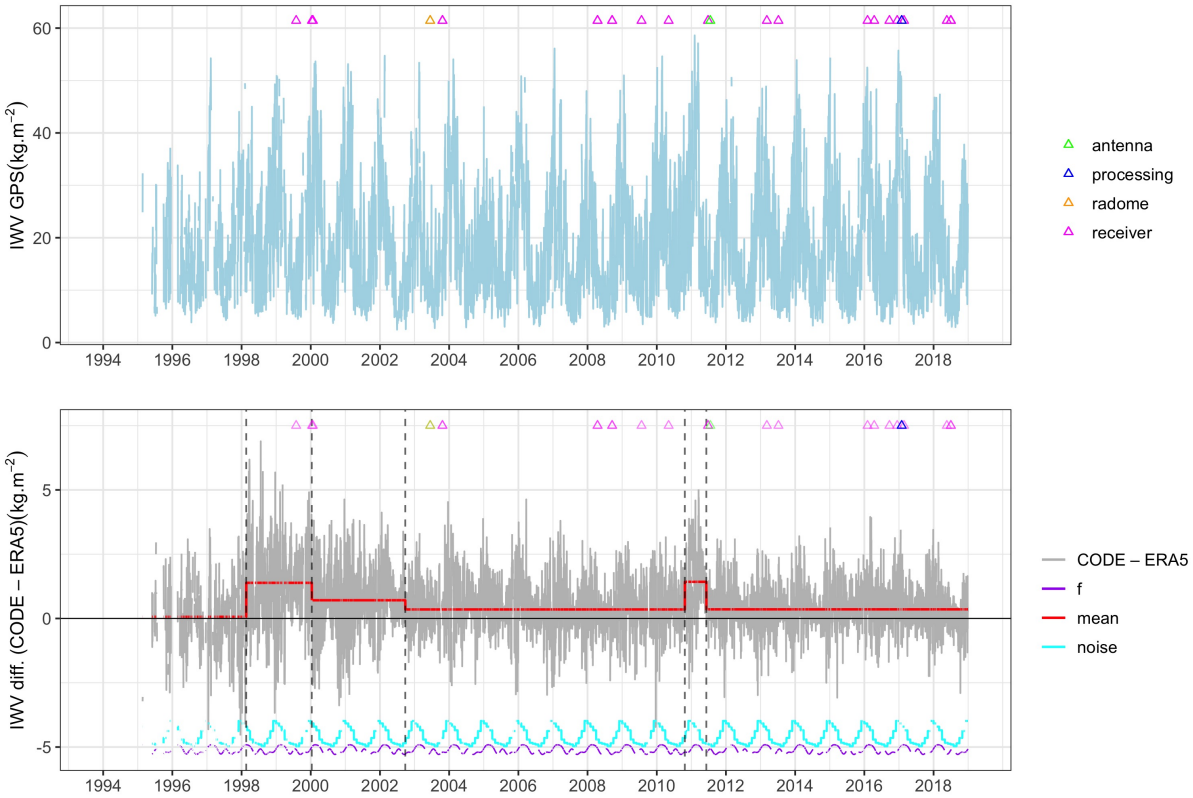


Figure 1.5 – Daily time series of IWV GNSS (top) and IWV difference between GNSS and ERA5 reanalysis (bottom) at station ALIC (Alice Springs, Australia). The bottom plot includes segmentation results superposed (the red line shows the mean for each segment, the vertical black dashed lines indicate the change-point positions, the cyan and purple lines at the bottom represent the estimated monthly mean variance and the functional modeling the periodic bias, respectively). Colored triangles on top of each plot indicate the position of known instrumental or processing changes from the GNSS metadata.

1.2 Review of existing homogenization methods

Homogenization refers to the process of identifying and correcting (artificial) inhomogeneities within data series. In this study, our focus is on the detection of abrupt changes or change-points within the data, as they can significantly impact the trend estimates. Two approaches have been commonly used for the detection step: either directly on the raw series, known as the "absolute" approach, or on the series of differences (as mentioned in the previous subsection), known as the "relative" approach. To mitigate the impact of large seasonal variations in the raw signal, the absolute approach needs either to include a mathematical model for the seasonal variations or to work on anomalies (differences with respect to the mean annual variation) instead of the raw signal. In the relative approach, this problem is greatly reduced but the introduction a reference series also introduces uncertainty on the origin of the detected change-points if the reference series is not perfectly homogeneous. Both approaches have been discussed and an exhaustive number of existing software packages have been assessed in the COST Action HOME (Venema et al., 2012). Their conclusion was unequivocal on the higher performance of the relative approach, although the attribution step remains a limiting factor.
In the rest of this section, we describe the different steps necessary to the relative homogenization approach.

Data preparation. The first task is to compute the series of differences between the target series and a reference series. The reference series can be taken from a single nearby station or computed as a composite from several nearby stations. It can also be taken from a different data source, such as reanalysis, as long as it shares a common climate signal with the target series (the goal is to suppress this signal by taking the difference of series). This step may actually consist of several sub-steps to minimize the contribution of various error sources present in the raw series or introduced in the series of differences, such as outliers and representativeness differences. The representativeness differences may usually recover two types of problems. The first one was already mentioned above for the case of the reanalysis. Indeed, reanalysis data are representative of aerial averages on the order of (twice) the product grid spacing (e.g. about 50 km for the 0.25° grid size of ERA5), while GNSS observations are more representative of the atmosphere in the close vicinity of the antenna, or a few km around it. The second type arises when a nearby station is used as a reference and there is a significant climatic difference between stations (e.g. when the distance is too far, e.g. hundreds of km, or even a shorter distances of a few km in mountainous areas). Another important feature is to make the IWV data consistent in the vertical dimension because water vapour density and thus IWV are quickly decreasing with altitude. Therefore, a correction of the IWV data due to vertical distance is required (Bock et al., 2022).

Segmentation. The central step of homogenization the segmentation which aim is to detect change-points (usually at unknown time). This can be done manually, by the visual inspection of time series, fully automatically, using a statistical method, or in a hybrid way where a statistical method is used to detect potential cnahge-points which are validated afterwards by an expert. A comprehensive review of segmentation methods used for the homogeneization of climate data has been done in Quarello (2020). The list of the current segmentation methods, along with associated software (if available), can be found in Table 2.1 of that thesis, as well as in references therein. Most of these methods have been assessed in the COST Action HOME (Venema et al., 2012), for monthly temperature and precipitation observations, and more recently in COST Action GNSS4WSWEC for GNSS IWV observations (Van Malderen et al., 2020). These segmentation methods are classified based on their model type (parametric or non-parametric), approach (frequentist or Bayesian), infer-

ence procedure (likelihood, penalized likelihood, or test), and search algorithm (optimal or sub-optimal), as shown in Figure 2.7 in Quarello (2020). Most of these methods operate using the relative approach, which is a common method for reducing strong seasonal variations and increasing the detection power of the segmentation.

In this study, we employ the segmentation method developed by Quarello et al. (2022) in a parametric and frequentist framework. This method allows to detect all change-points simultaneously in an efficient manner or using an optimal search computationally through the use of dynamic programming. This method has been specifically developed for detecting change-points in Integrated Water Vapour (IWV) difference series between GNSS and reanalysis data. In particular, a change-point detection in the mean model is proposed taking into account the characteristics of the target series such as periodic bias and varying variance as shown in the example in Figure 1.5. The inference of the model is done through a two-step strategy: first, for a given number of change-points, the parameters of the model (the means of the segments, the variances of each month, the periodic function and the change-points) are estimated using the classical maximum likelihood procedure. Second, the number of change-points is estimated. This segmentation method effectively addresses two fundamental questions in the procedure, an algorithmical and a statistical issues:

- ⋆ Estimation of change-points position: the classical dynamic programming (DP) algorithm is used to locate change-points. This algorithm is now well known to retrieve the change-point efficiently (fastly and exactly) computationally speaking. This algorithm can be used if and only if the quantity to be optimized is additive with respect to the segment (see details in sub-section 2.2.1). The proposed inference method has been developed in order to be able to use it.
- ⋆ Estimation of the number of change-points: this issue is resolved using penalized criteria, as classically used in segmentation frameworks. A penalty term is added to the inference criterion (e.g., log-likelihood) to account for the model's complexity, i.e. the number parameter to be estimated.

Quarello (2020) extensively evaluated this segmentation method by comparing its performance using four penalized criteria, in both simulation and real data scenarios. The final version was released as the R package named "GNSSseg" and published (Quarello et al., 2022). Furthermore, this method underwent a comparison with other segmentation methods in a benchmark exercice coordinated by Van Malderen et al. (2014) where it demonstrates strong potential and revealed as one of the best methods.

Attribution/validation. Change-points identified through segmentation can originate from either the target or the reference series. Therefore, attributing the detections to the right series is crucial before correcting the raw IWV data for the jumps. There are two primary strategies to address this attribution problem. The first approach aims to improve the homogeneity of the reference series, thus making it possible to attribute all change-points to the target series. Typically, these reference series are generated by averaging data from multiple nearby stations (Alexandersson, 1986; Menne and Williams, 2005; Thorne and Vose, 2010). However, in practice, even the composited series often have non-negligible inhomogeneities. The second approach looks for the origin of the change-point through pairwise comparisons. This can be achieved semi-automatically, employing statistical inference and manual cross-checking with historical information (station metadata), in an iterative way (Caussinus and Mestre, 2004). However, there could be gaps or mistakes in the available metadata. Another method, proposed by Menne and Williams (2009), fully automates the process by assigning change-points to the station that has the most detections, when segmenting all possible combination series of

differences. This method therefore requires a large number of nearby stations to be effective. This approach also leverages station metadata when available and attributes detected change-points to the closest known events in the station's history within specified confidence limits.

Correction. The final step involves correcting the target time series for the jumps in mean corresponding to the attributed or validated change-points only. The traditional way of doing this is to subtract the estimated offsets from the GNSS IWV series while leaving the most recent homogeneous segment unchanged (the rationale for this is that the most recent data may have smaller absolute bias than the older data) (Van Malderen et al., 2020).

1.3 Objectives and outline of this thesis

The objective of this study is to establish a comprehensive homogenization procedure for the IWV time series derived from GNSS data using the segmentation method developed by Quarello (2020). Therefore, this thesis focuses on two major tasks:

- ⋆ The first task involves a rigorous investigation of the robustness, sensitivity, and inherent limitations of the segmentation method. This endeavor is particularly crucial for practical real-world applications. The scientific questioning and factors influencing this analysis encompass the quality of the GNSS data which primarily depend on the GNSS data processing methodologies (do more recent GNSS solutions have reduced inhomogeneties?), the length of the analysed time series (as the length is increasing with years, what is the impact on the segmentation performance?), variations in auxiliary data employed in IWV conversion, and the selection of the data source for the reference time series (multiple choices can be made among various reanalysis products). Additionally, the study explores the extent to which changes in the segmentation results impact the subsequent IWV trend estimates (is it more beneficial to detect and correct more small jumps or fewer but larger ones?).
- ⋆ The second task centers on the development of an automatic attribution method. The method is intended to operate as a post-processing step separate from segmentation and should be applicable even in cases of sparse GNSS networks where only a few nearby station are within a reasonable (100-200 km) distance.

The thesis is structured as follows: Chapter 2 provides an introduction to the datasets employed and to the fundamental mathematical concepts and tools utilized in this study. This chapter covers: i) the retrieval of IWV from GNSS observations, including the basic principles of GNSS data processing; ii) a brief description of the mathematics of the GNSSseg segmentation method developed by Quarello (2020); iii) fundamentals of linear regression methods and linear stochastic model identification, with a special focus on the mixed autoregressive - moving average (ARMA) models that will be effectively utilized for modeling the noise in the IWV difference series; and iv) an introduction to a number of classification algorithms employed in the development of the attribution method. Chapter 3 is dedicated to the first task. It is presented as a paper published in the Atmosphere journal. In this paper, it is important to note that the segmentation results and trend estimates did not yet include the attribution step which was developed afterwards. Especially, for the trend estimates, the results include two versions: one with all detected change-points corrected (as if only GNSS data were inhomogeneous) and one with only the change-points that could be directly validated with the help of metadata. The main caveat with the latter is that changes due to external factors, not reported in the metadata, are not taken into account. Some true change-points may thus remain in the corrected time series. Chapter 4 comprises two main sections.

First is a paper, submitted to the International Journal of Climatology (presently under review) which described and applies the developed attribution method on a real data set of 81 main stations. Second in this chapter is additional research conducted alongside the paper's topic. Finally, Chapter 5 summarizes the main findings presented in this thesis and discusses future perspectives. The thesis contains also an Appendix section which presents some preliminary results of the application of the full homogenization process to an extended GNSS dataset comprising more than 6000 stations.

Chapter 2

Data and mathematical tools

2.1 Data

2.1.1 GNSS data

GNSS Integrated Water Vapor (IWV) derives from the propagation delay of radio waves in the atmosphere on their path from the GNSS satellites to the ground-based receivers (Figure 2.1). Data processing with scientific software and use of a number of auxiliary geodetic products is required to achieve estimates of propagation delays and subsequent IWV data with accuracy compatible for meteorology and climatology. In this sub-section we summarize the main steps and highlight the crucial aspects of data processing intervening in the elaboration of GNSS IWV data.

2.1.1.1 GNSS observations and data processing methods

GNSS are satellite radio navigation systems providing global positioning everywhere on Earth (Hofmann-Wellenhof et al., 2007). Their principle is based on the transmission of radio signals by satellite constellations which allow receiver-equipped users to access precise three-dimensional positioning in real time with precision 1-10 m. The most popular GNSS is the US GPS (Global Positioning System), which is fully operational since 1994. It has been complemented over the years by the Russian GLONASS, the European Galileo, and the chinese Beidou, among others (Teunissen and Montenbruck, 2017a, part B). Ground segments, relying on tracking networks and data processing facilities, are associated to the space segment to provide essential navigation products such as satellite ephemerides, clock synchronization information, and Earth orientation parameters (the latter are necessary to relate the the inertial reference frame, which is natural to satellite orbits, to the Earth-fixed frame used to express the receiver coordinates). Highly accurate versions of these products are computed in near real time by the International GNSS Service (IGS) and International Earth Rotation and Reference Systems Service (IERS) analysis centres for scientific usage.

Some of the inhomogeneities that we are tracking in the GNSS IWV time series can be related to changes in the observations, data processing procedure, and auxiliary products used for the processing and/or the conversion. To understand how these features can impact the propagation delays and IWV, it is necessary to provide some insight into the data processing procedure.

Geodetic positioning relies on the analysis of the so-called "phase observations". A single phase observation between a receiver r and satellite s can be formalized by the following equation (Bock, 2012, p. 6):

$$L_r^s = \rho_r^s + c(\Delta t_r - \Delta t^s) + \lambda N_r^s + \Delta\rho_{rel} - \Delta\rho_{iono} + \Delta\rho_{tropo} + \Delta\rho_{ant}^s + \Delta\rho_{ant,r} + \epsilon. \tag{2.1}$$

where:

- ⋆ ρ_r^s is the geometric distance between the receiver's antenna reference point (ARP) and satellite ARP,
- ⋆ c is the speed of light in a vacuum,
- ⋆ Δt_r and Δt^s are the receiver and satellite clock offsets from the GNSS time reference,
- ⋆ λ is the wavelength of the radio signal,
- ⋆ N_r^s is the integer phase ambiguity,
- ⋆ $\Delta\rho_{rel}$ account for relativistic effects,
- ⋆ $\Delta\rho_{iono}$ is the propagation delay in the ionopshere,
- ⋆ $\Delta\rho_{tropo}$ is the propagation delay in the troposphere,
- ⋆ $\Delta\rho_{ant}^s$ and $\Delta\rho_{ant,r}$ are the phase delays in the satellite and the receiver antennas,
- ⋆ ϵ accounts for other minor errors.

Basically, GNSS data processing is a regression problem in which unknown parameters of the observation equation are estimated in order to minimize the errors between the observed and predicted phase measurements. The IGS standards in "static" geodetic processing (i.e. positioning of ground-fixed antennas) consists in estimating the parameters in (2.1) in 24 hour batches or "sessions". Observations are typically sampled at a 30-sec rate, while the sampling of parameters is much lower (e.g. 1 set of receiver coordinates per session, 1 hour sampling for tropospheric delay parameters, etc.). However, not all parameters can be estimated simultaneously because of multi-colinearity. Namely, satellite positions and clock delays are usually fixed to their values computed by the IGS (so-called IGS satellite products), and relativistic effects which can be predicted by theory with a high degree of precision (Zhu and Groten, 1988) are corrected a priori. Receiver clock offsets can be either estimated "by epoch" (one parameter per observation) or eliminated by combining the simultaneous observations from two satellites and two receivers (so-called double differences) but they require that the data from multiple stations are used once. The ionospheric delays are usually eliminated by combining instantaneous observations from 2 or 3 frequencies, thanks to the dispersive nature of the refraction index of the ionospheric plasma region at the GNSS frequencies.

Phase delays in the satellite and receiver antennas are an important source of bias that needs be corrected (Hofmann-Wellenhof et al., 2007). Empirical models are used therefore, comprising a constant offset component and a variable hemispheric model which accounts for variations depending on the angle of incidence and frequency of the radio signal with respect to the antenna's orientation. These variations are contingent on the specific antenna model and type in use. For the satellite, the antenna offset is expressed relative to its center of mass and is on the order of 1 m. The offsets for the receiver antennas are given relative to the ARPs and are of order 0.1 meter. The variations are typically one order of magnitude smaller but need also to be corrected.

Eventually, station positions, receiver clock offsets, tropospheric delay, and ambiguity numbers are the main parameters that are estimated.

When the observations are close to zenith, several of these parameters are tightly correlated (station height, receiver clock offset, tropospheric delay, and ambiguity number) and biases in the a priori correction models (e.g. antenna phase model offsets) and/or in the auxiliary products can map onto all these parameters (Teunissen and Montenbruck, 2017b). One way to reduce this problem is to include observations at low elevation angles.

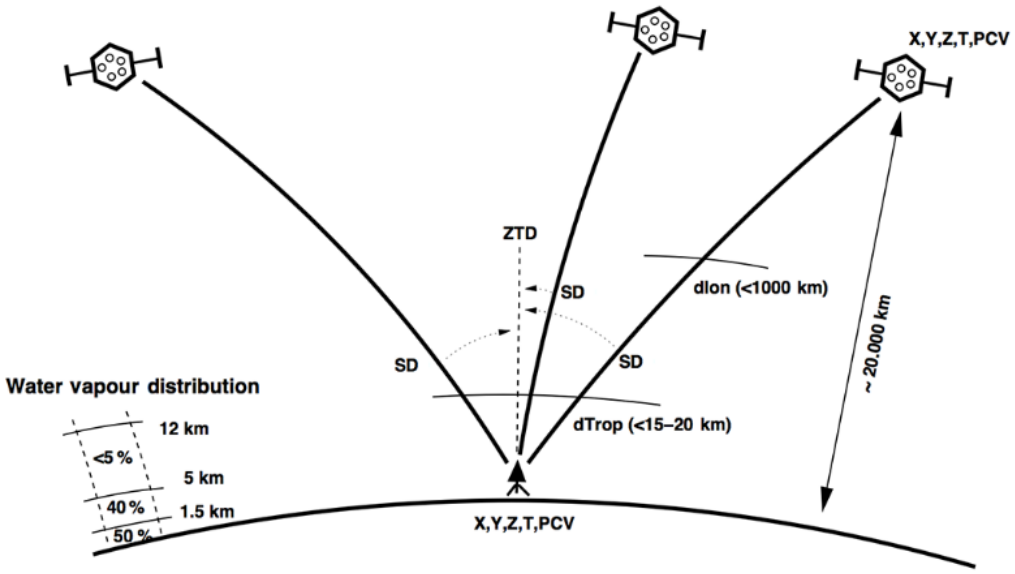


Figure 2.1 – Schematic presentation of individual slant delays (SD) from GNSS satellites and their mapping to zenith tropospheric delay (ZTD), (Guerova et al., 2016).

Apart from the factors explicitly introduced in equation (2.1), the error term, ϵ, includes all other error sources, namely the "multipath" effect, which is a spurious phase modulation induced by reflective surfaces in the vicinity of the receiving antenna. Changes in the reflective properties of the environment around the receiver's antenna can modify this effect and be a source of inhomogeneities in the estimated parameters (positions, tropospheric delays, etc.).

2.1.1.2 Modelling the tropospheric delays

GNSS signals undergo both retardation and bending as they cross the troposphere and the lower stratosphere, because of the varying index of refraction along their propagation path. The combined effect is referred to as the "total tropopsheric delay". It is also sometimes called "excess path length", in reference to geometrical optics, and denoted ΔL, but it is identical to the term $\Delta\rho_{tropo}$ introduced in equation (2.1). Below we use the traditional notation, following Bevis et al. (1992), which writes:

$$\Delta L = \int_L n(s)ds - G, \tag{2.2}$$

where $n(s)$ corresponds to the index of refraction of the air at the position s along the curved trajectory of the ray path denoted as L. The parameter G signifies the straight-line geometrical path length through the atmosphere

assuming $n = 1$. A modified form of Equation (2.2) writes:

$$\Delta L = \int_L [n(s) - 1] ds + S - G.$$

where, $S = \int_L ds$, is the curved path length along L. This equation provides a breakdown of the excess path length, with the first term on the right side accounting for the retardation effect (reduction the wave velocity due to $n > 1$), ΔL_r, and the second term representing the bending effect (a purely geometrical effect). The bending effect is much smaller than the retardation at most elevation angles above 0°. Negelecting the bending effect, the first term is often re-expressed in terms of atmospheric refractivity $N = 10^6 \times (n - 1)$ as follows (Bevis et al., 1992):

$$\Delta L = 10^{-6} \int_L N(s) ds \tag{2.3}$$

The refractivity of the air can be decomposed into two basic groups of constituents, dry air and water vapor molecules. Following (Davis et al., 1985), the tropospheric delay is reformulated to account for this distinction:

$$\Delta L = 10^{-6} \int_L N_h(s) ds + 10^{-6} \int_L N_w(s) ds = \Delta L_h + \Delta L_w. \tag{2.4}$$

where the first integral is referred to the "hydrostatic delay", primarily due to the delay induced by dry air, and the "wet delay", due to the water molecules. Furthermore, Saastamoinen (1973) first noted that the hydrostatic delay integral taken in the zenith direction is simply proportional to surface pressure (by use of the hydrostatic equation in the atmospheric column). This property offers a strong constraint on the determination of the total delay in Equation (2.1) as the Zenith Hydrostatic Delay (ZHD) can be corrected a priori from auxiliary surface pressure data (from measurements or atmospheric model outputs), leaving only the Zenith Wet Delay (ZWD) as unknown parameter. Because the phase observations are collected in practice from satellites in any direction (see Figure 2.1), mapping functions are used to project the zenith delays onto the slant paths and vice versa.

Current state of the art GNSS processing software use slightly more sophisticated tropospheric models where, in addition to the ZHD and ZWD parameters, horizontal gradients are included to describe a first order azimuthal variation in the tropospheric refractivity field above the antenna (Davis et al., 1993):

$$\Delta L = \Delta L_h^z \times m_h(\epsilon) + \Delta L_w^z \times m_w(\epsilon) + (\vec{G} \cdot \vec{e}) \times m_G(\epsilon) \tag{2.5}$$

where ΔL_h^z and ΔL_w^z are the zenith hydrostatic and wet delays, respectively, and $\vec{G}$ is the horizontal gradient vector, $\vec{e}$ signifies the unit vector pointing in the direction of the observed satellite, and m_h, m_w, and m_G are the mapping functions for the ZHD, ZWD, horizontal gradient, respectively. Here ϵ represents the elevation angle.

In summary, the handling of tropospheric delay in GNSS data processing consists in mainly specifying: i) the mapping functions and the source of surface pressure or ZHD data (e.g. from an atmospheric model), ii) the rate at which ZWD and gradient parameters should be estimated (e.g. once per hour) to account for atmospheric variability, iii) the constraints on the temporal variation of the estimated parameters.

2.1.1.3 Conversion of tropospheric delay to IWV

The GNSS IWV product is derived in a post-processing step which consists in extracting the ZWD part from the Zenith Tropospheric Delay (ZTD) that is the standard output of the processing software. For this purpose, we don't use the ZHD that served for the a priori correction in Equation (2.1) but a more accurate estimate if one is available. The reason is that with the current state of the art GNSS processing standards, the prior ZHD is taken from a coarse resolution atmospheric model (e.g. $2° \times 2.5°$, see next sub-section), which is also used to compute the mapping function models. A more accurate ZHD estimate can be computed from a high-resolution version of the atmospheric model. In this work we use ERA-Interim or ERA5, which have a grid resolution of 0.75° and 0.25°, respectively, in both dimensions. Once an accurate estimate of ZHD is subtracted from the ZTD, the ZWD component needs to be converted into IWV by application a "delay to mass" conversion factor, denoted κ. The definition of κ stems from the ratio of IWV defined by Equation (1.3) and ZWD defined by:

$$\Delta L_w^z = 10^{-6} R_v \int_0^\infty \rho_v(z) \times (k_2' + \frac{k_3}{T_m}) dz \tag{2.6}$$

where k_2' and k_3 are refractivity coefficients for the water molecule, and R_v is the specific gas constant for water vapor. Various empirical values for the refractivity coefficients are available (Bevis et al., 1994). In this work we used the updated version proposed by Bock (2020a). Forming the ratio of Equations (1.3) and (2.6), we get:

$$\kappa(T_m) = \frac{10^6}{R_v(k_2' + \frac{k_3}{T_m})}$$

where T_m, referred to as the "weighted mean temperature", is given by:

$$T_m = \frac{\int \rho_v(z) dz}{\int \frac{\rho_v(z)}{T(z)} dz} \tag{2.7}$$

The conversion of ZWD into IWV finally writes:

$$IWV = \kappa(T_m) \times ZWD. \tag{2.8}$$

A rule of thumb for Equation (2.8) is $IWV = 155 \times ZWD$ when ZWD is expressed in m, or $IWV = ZwD/6.5$ when ZWD is expressed in mm, which means that 6.5 mm of wet delay at zenith converts to 1 kg m^{-2} of integrated water vapour (Bock, 2020a). This rule is useful to quickly convert delay errors into IWV errors.

For a more accurate conversion of estimated ZWD data, knowledge of the T_m parameter is necessary. Examination of (2.7) reveals that, although both the vertical profile of water vapour density, $\rho_v(z)$, and temperature, $T(z)$, appear in the definition of T_m, the former is both in the numerator and denominator of the fraction, and therefore the computed T_m is only sensitive to second order to errors in $\rho_v(z)$. This suggests that T_m could be parameterized as a function of temperature. Bevis et al. (1992) showed that T_m can be predicted to a good accuracy from surface air temperature, and proposed an empirical model obtained by linear regression. Such a model is, however, highly site dependent, and needs to be recomputed for stations in different climatic regions. Moreover, the prediction based on surface temperature tend to map spurious diurnal variations into T_m. Because of these limitations, it is preferable and technically more convenient to use T_m estimates based on a high resolution atmospheric model (Bock, 2020a).

In this work, the ZHD and T_m conversion parameters were both computed from ERA-Interim or ERA5 vertical profiles at their nominal horizontal grid resolution. The uncertainty in these parameters is assumed to be below 2.5 mm in ZHD (equivalent to an error in surface pressure of 1 hPa) and below 3 K in T_m. The induced IWV equivalent uncertainty is 0.4 kg m^{-2} and 0.3 kg m^{-2}, respectively, assuming a typical mid-latitude T_m of 275 K and IWV of 25 kg m^{-2}.

2.1.1.4 Important data processing aspects

GNSS processing can be carried out using various software packages and processing options. In this subsection, we explore the significant factors influencing the estimated ZTD accuracy, such as processing mode, tropospheric modeling aspects, session duration, elevation cutoff angle, elevation-dependent observation weighting, and correction models for antenna phase center offsets and variations.

Scientific GNSS software packages are generally based on one of two processing modes: double difference (DD) positioning, which uses double-differenced phase observations from a network of stations, or precise point positioning, so-called PPP, which use the undifferenced phase observations (Teunissen and Montenbruck, 2017a, p. 13). DD processing is independent of external satellite clock products, while PPP relies on the precise satellite orbits and clocks. The increased availability of precise orbit and clock solutions in the late 1990s, thanks to the efforts of the IGS, has been instrumental. Moreover, PPP demands precise models to correct for systematic effects causing centimeter-level variations (e.g. the Higher-Order Ionospheric Delay Corrections, Site Displacement Effects, etc). An overview of the various model components and corrections in PPP applications is provided in Teunissen and Montenbruck (2017a, p. 727).

Several aspects of tropospheric modelling have been implemented differently depending on the software. This concerns especially the modelling of the temporal variations of the estimated ZWD and gradient parameters. For example, in the Bernese software, a piece-wise linear model is fitted using the least squares method (Dach et al., 2015). Conversely, the GIPSY software utilizes a random walk model (Zumberge et al., 1997), and the GAMIT software relies on the Gauss-Markov model (Herring et al., 2015). While GAMIT is also based on a least-squares estimation, the GIPSY software implements the Kalman Filtering technique. The latter approach makes it possible to estimate tropospheric parameters at each epoch of observations (e.g. every 5 min), but this requires to set adequate constraints in the random walk model.

Mapping functions are crucial elements of the tropospheric model as they represent the partial derivatives of the estimated parameters in Equation (2.5). Currently, the two widely used mapping functions are the first Vienna mapping function (VMF1) (Boehm et al., 2006b) and the Global Mapping Function (GMF) (Boehm et al., 2006a). They are both built on the concept of the Niell Mapping Function (NMF) (Niell, 1996) and the Isobaric mapping Function (IMF) (Niell, 2000) but use numerical weather model data which provide a more comprehensive description of the refractive index in space and time. These mapping functions are parameterized with a number of coefficients that need are extracted at the specific GNSS station locations by the processing software. In the case of the VMF1, the coefficients are provided with a 6-hourly time step from the operational ECMWF analyses on a 2° latitude by 2.5° longitude horizontal grid. This mapping function, along with its

more recent version VMF3, are the most accurate. Currently, VMF1 is recommended by the IERS Service.

Another important feature is the correction of the antenna phase-center offsets (PCOs) and variations (PCVs), both on the satellite and the receiving side (Teunissen and Montenbruck, 2017a, Chap. 17). The magnitude of the PCVs depends on the elevation angle of the satellite as observed from the ground, and it can vary between different types of antennas. These offsets and variations have a significant impact on the estimated ZTD delay. Use of a proper correction models is essential to address this issue. Earlier models (before 2006) were based on a relative calibration technique and introduced a systematic bias in the absolute ZTD estimates. The more recent, absolute calibrations, have resolved this problem and have been adopted since 2006 in the IGS. However, for a small number of old antennas, used in the IGS network and elsewhere, only relative calibration models are available. This issue is discussed in more detail in Section 2 of paper 1 for the subset of IGS stations used in our work.

ZTD and several other parameters estimated during the GNSS processing are sensitive to elevation dependent errors. Such errors may especially arise from errors in the mapping functions which represent a smooth relationship between zenith delay and slant delay, whereas the true atmosphere may be more complex in many situations (e.g. a passing meteorological frontal system). They may also be due to the application of a wrong or an imperfect antenna PCO/PCV model, as well as as to multipath and signal interference (Teunissen and Montenbruck, 2017a, Chap. 15 and 16). Mapping function errors and multipath errors are typically larger at lower elevations. One commonly used strategy to minimize their impact is to increase the elevation cutoff angle or to down-weight low-elevation observations. However, low-elevation observations improve geometry and reduce formal ZTD errors. Thus, an optimum combination of elevation cutoff angle and weighting strategy needs to be adjusted to mitigate these errors.

The overall uncertainty in the ZTD and ZWD estimates resulting from the typical current state-of-the art GNSS processing strategies is estimated to 2-5 mm (Teunissen and Montenbruck, 2017a, Chap. 38). Using the rule of thumb introduced in previous sub-section, the ZTD uncertainty adds a further 0.3 to 0.8 kg m^{-2} uncertainty to the derived IWV estimates.

2.1.1.5 GNSS tropospheric delay products used in this work

In this thesis, we used three distinct GNSS datasets, each representing a different generation of (re-)processing products: IGS repro1 (first generation), CODE REPRO2015 (second generation), and NGL repro3 (third generation).

The IGS repro1 dataset was produced between 2010 and 2011 by JPL/NASA in the framework of the first reprocessing campaign organized by IGS (Byun and Bar-Sever, 2009). It marks the first comprehensive reprocessing effort with collaborative contributions from multiple Analysis Centers. This reanalysis encompassed the entire history of GPS data collected by the IGS global network from January 1, 1995, to December 31, 2007, employing the best available models and methodologies at the time. JPL completed this dataset consistently by mid-2011, providing tropospheric products (ZTD and gradients)for 460 IGS stations, for the period 1995-2010 (inclusive). A quality-checked and IWV-converted sub-set of 120 stations with more than 15 years of measurements was prepared by Bock (2016) and used by Parracho (2017) and Quarello (2020).

The second dataset, CODE REPRO2015, was processed by the Center for Orbit Determination in Europe (CODE) in 2015, using the framework settled for second IGS reprocessing campaign (Dach et al., 2015). It includes 434 IGS stations and covers the period from January 1, 1994, to December 31, 2014. Until now, this dataset has been extended year after year with the operational CODE product (Dach et al., 2018), and quality-checked and IWV-converted estimates have been released by Bock (2019) .

Details of the processing methodologies for the IGS repro1 and CODE REPRO2015 datasets can be found in section 2 of paper No. 1 where various important features described in the previous sub-section are summarized and discussed. To mention just one, we note that IGS repro1 was produced with GIPSY OASIS II software in PPP mode and CODE REPRO2015 with the Bernese GNSS software in a DD positioning mode. For the purpose of paper No. 1, we selected 81 common stations, ensuring that the time series in both datasets covered a minimum period of 15 years. The map in Figure 1 in this paper displays the available stations from both datasets and highlights the selected stations.

Software	GipsyX Version 1.0
Strategy	Precise Point Positioning (PPP)
Orbits, Clocks, ERPs	Daily Repro3.0 (JPL)
Reference Frame	IGS14
Antenna Calibration	igs14_www.atx
Window Length	30 hours
Elevation Cutoff Angle	7°
Observations	GPS
Observation Sampling Interval	5 minutes
Observation Weighting	$\sigma^2 = 1/sin(e)$ where e is the elevation
Tropospheric model	ZHD and ZWD a priori: 6-hourly ECMWF analysis (provided by TUV) VMF1 mapping functions (hydrostatic and wet). Random Walk model for ZWD and gradient parameters with constraints: 3 mm $h^{-1/2}$ (ZWD) and 0.3 mm $h^{-1/2}$ (gradients) ZWD and gradient sampling: 5 min
Tropo files	ZTD and gradient estimates provided in SINEX files (0000, 0005, ...2345 UTC)
Coordinate estimates	Estimated once per 24 h
Ambiguity resolution	Fixed

Table 2.1 – Summary of GNSS data processing parameters of the NGL repro3 dataset.

The third dataset, NGL repro3, is sourced from the Nevada Geodetic Laboratory (NGL) (Blewitt et al., 2018). It represents the latest generation of GNSS products. NGL routinely collects and processes geodetic-quality

GPS observations from more than 20,000 stations worldwide, encompassing various regional and commercial networks, in addition to the widely used IGS network. The dataset covers the period from 1994 to present. Details of the processing relevant to ZTD estimation are provided in Table 2.1. One can note that NGL uses the GipsyX version 1.0 software from JPL in PPP mode together with JPL's repro3 final GPS orbits and clocks. The IWV data from this dataset were derived by applying a quality-checking and conversion procedure consistent with the other two datasets. The daily IWV estimates were used for the attribution task in the second paper (Chapter 4). In this work, the above-mentioned 81 stations from the CODE REPRO2015 dataset were selected as the main stations together with 628 nearby stations from the NGL dataset.

The NGL repro3 dataset uses a more modern version of JPL's GNSS processing software, and thus shares many similar features with the IGS repro1 dataset, including PPP mode, ZWD and gradient stochastic model, and elevation cutoff angles, among others. Its mapping functions are more modern, and in line with CODE REPRO2015. In contrast to IGS repro1, NGL repro3 adopted a strategy that down-weights low-elevation observations, making it less sensitive to multipath errors. The fixed ambiguity resolution approach is also an efficient way to reduce ZTD biases. NGL repro3 adopted also the more recent reference frame, IGS14 and the associated igs14.atx. The latter has resulted in a significant increase, up to 90%, in the number of stations within the IGS network that have now absolute calibrations (Rebischung and Schmid, 2016).

2.1.2 Reanalysis data

Climate reanalysis combines modern weather forecasting models with historical observations through data assimilation to delivers a global picture of weather and climate of the past as close to the reality as possible. This representation is known for its global coverage and temporal consistency, often described as "maps without gaps", extending from the Earth's surface to the top of the atmosphere (*Fact Sheet: Reanalysis* 2023). They utilize realistic models to extrapolate information from locally observed parameters to nearby locations and forward in time. Unlike operational analyses produced by Numerical Weather Prediction systems, reanalyses are created using a consistent assimilation system and the same numerical model, making them unaffected by changes in methods, model physics, or dynamics.

In our research, we focused on two specific reanalysis datasets produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) for climate mornitoring: ERA-Interim (Dee et al., 2011) and ERA5 (Hersbach et al., 2020). ERA-Interim is ECMWF's previous atmospheric reanalysis, based on a 2006 version of the Integrated Forecasting System (IFS) physical model. Its data span from 1979 until August 31, 2019, and no further updates are available. ERA5, on the other hand, represents the latest (Fifth) generation of ECMWF global reanalysis, based on a 2016 version of the ECMWF IFS. ERA5 offers data starting from 1959 continuously expanded in near real-time. ERA5 offers higher spatial and vertical resolutions (0.25° and 137 levels) compared to ERA-Interim (0.75° and 60 levels). Additionally, ERA5 benefits from increased satellite observations and improvements in its operational processes. As an example, ERA5 assimilated several satellite channels sensitive to humidity through an all-sky assimilation approach, in contrast to ERA-Interim, which employs a clear-sky assimilation approach. Moreover, ERA5 uses a 10-member ensemble 4D-Var system, providing estimates of uncertainties in the data, a feature not present in ERA-Interim (Hersbach, 2019). Various enhancements are expected to yield more accurate IWV estimates, as well as improvements in surface pressure

and air temperature, which are used as auxiliary data for IWV conversion.

We anticipate that the representativeness difference (Bock and Parracho, 2019) between GNSS and ERA5 will be smaller compared to that between GNSS and ERA-Interim, primarily due to ERA5's improved spatial resolution. This difference stems from the fact that the reanalysis grid-point values are representative of aerial averages while the GNSS observations are comparatively equivalent to point observations. It is important to note that we include ERA-Interim primarily because this work builds on the previous work by Parracho (2017) and Quarello (2020), and we wanted to investigate the impact of using different reanalysis data in the homogenization process.

It is essential to emphasize that the homogeneity of reanalysis data is uncertain. As mentioned earlier, ERA-Interim showed inhomogeneities when changes occurred in the observing system (Schroeder et al., 2016), and perhaps ERA5 as well (Allan et al., 2022). For these reason, we also consider the possibility of having change-points in the reanalysis used as a reference in forming the IWV difference series in the homogenization process.

2.2 Mathematical tools

2.2.1 GNSSseg segmentation method

In this section we recall briefly the segmentation method developed by Quarello et al. (2022) for the change-point detection problem in GNSS series (being the basis of the work of this thesis).

2.2.1.1 Model

Quarello et al. (2022) modeled the series of IWV difference by a Gaussian independent random process $z = \{z_t\}_{t=1,...,n}$, such that

$$z_t \sim \mathcal{N}(\mu_k + f_t, \sigma^2_{month}) \text{ if } t \in I^{mean}_k \cap I^{var}_{month} \text{ for } k = 1, ...K,$$

where

- ⋆ z_t is the data at time t,
- ⋆ K denotes the number of segments,
- ⋆ μ_k represents the constant mean of the kth segment $I^{mean}_k = [[t_{k-1} + 1, t_k]]$, with t_k indicating the position of the kth change-point with the convention $t_0 = 0$ and $t_K = n$,
- ⋆ f_t a function of time modeled by a Fourier series of order 4 accounting for the periodic bias:

$$f_t = \sum_{i=1}^{4} a_i cos(2\pi t/T) + b_i sin(2\pi t/T),$$

  with T = 365.25 (days) and the coefficients a_i and b_i are weights for the series representing yearly, semiannual, ter-annual, and quarterly cycles, corresponding to i ranging from 1 to 4 respectively,
- ⋆ σ^2_{month} stands for the monthly variance, consistent within the monthly interval $I^{var}_{month} = \{t, date(t) \in month\}$.

2.2.1.2 Estimation procedure

The proposed inference procedure is illustrated in Figure 2.2. This procedure consists of three main steps:

1. Estimation of the monthly variance σ^2_{month} as proposed by Bock et al. (2019).
2. For a fixed number of segments K, the periodic bias function f_t and both the change-points t_k and the segment means μ_k for $k = 1, ..K$, are iteratively estimated by minimizing the residual sum of squares, denoted by SSR_K:
$$SSR_K = \sum_{k=1}^{K} \sum_{month} \sum_{t \in I_k^{mean} \cap I_{month}^{var}} \frac{(z_t - f_t - \mu_k)^2}{\hat{\sigma}^2_{month}}.$$
This iterative procedure stops when the relative change in the estimated values is smaller than a threshold.
3. This estimation procedure is performed for different values of K (from 1 to $K_{\max}$ that is the maximal considered number of segments) and thus returns $K_{\max}$ different segmentations. The "best" one, i.e. the "best" number of segments is determined as follows:
$$\hat{K} = argmin_K SSR_K + pen(K).$$
Different penalty functions (pen) have been considered and tested in Quarello et al. (2022): mBIC (Zhang and Siegmund, 2007), Lav (Lavielle, 2005), BM1 (Birgé and Massart, 2001) and BM2 (Lebarbier, 2005). The idea of adding a penalty to SSR_K is to prevent overfitting, where the model attempts to have as many change-points as possible.

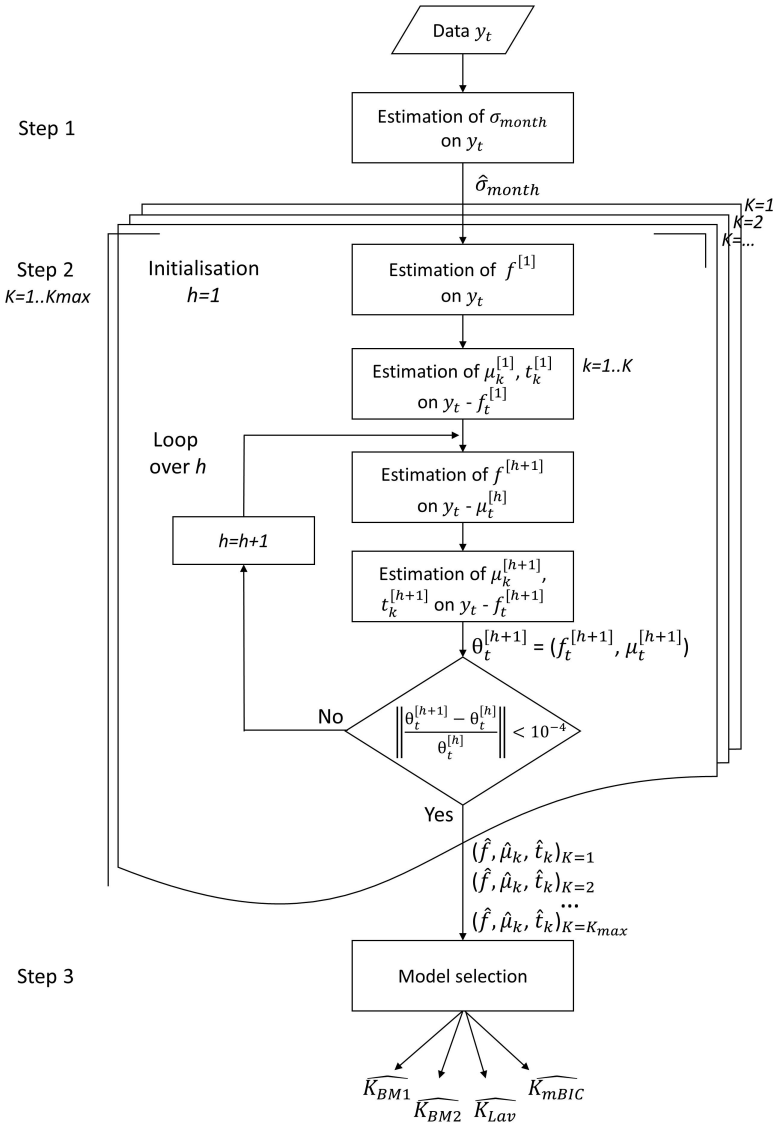


Figure 2.2 – Segmentation inference procedure (Quarello, 2020).

2.2.1.3 Important features

Certain factors need to be considered when employing this method.

The foremost factor concerns the selection of the number of change-points, K. Specifically, the challenge lies in selecting the appropriate criterion. Quarello et al. (2022) tested the four different penalized criteria through both simplified simulations and real data analysis (IGS solution). The simulation findings revealed that Lav exhibited instability, showing a large variance in detection numbers, whereas other criteria were more conservative. When applied to the real data, mBIC tended to over-segment compared to the other criteria. This tendency was consistent when we applied this method to other solutions (CODE). To select the final change-points, the authors proposed a semi-automatic approach which combines a visual inspection and the results from criteria. This approach involved inspecting the monthly time series for each station, prioritizing

based on known changes in GNSS metadata and information from the TEQC software. Through this validation process, Quarello et al. (2022) concluded that BM1 had the highest acceptance rate at 57%, followed by Lav at 51% and BM2 at 45%. Consequently, in our study, we focus solely on the results derived from the BM1 criterion.

Second, an important consideration is that no dependence between the data is assumed in this segmentation model. However, we have observed a time-dependent aspect within our data, a phenomenon commonly encountered in climate data as well (see the subsection 4.1.3.3). Neglecting this feature could lead to an oversegmentation of the series, as discussed in Chakar et al. (2017).

We acknowledge the potential issue of confusion between the estimates of f_t and μ_k particularly when dealing with seasonal variations. These variations can occasionally be mistakenly attributed to the mean rather than the periodic function, a situation that can arise with oversegmentation. Additionally, the absence of interannual variation within the model may lead to a false change in the mean to take it into account.

Finally, the method has been implemented in a R package named GNSSseg available on the CRAN, followed by a faster version called GNSSfast and available on github. In our study, segmentation results were obtained using the GNSSfast version. We adhered to all default settings, which include a maximum segment number of $K_{\max} = 30$ and an iterative estimation threshold set at 10^{-1} (the stopping rule of the iterative estimation procedure), etc...

2.2.2 Linear Regression

In this thesis, we considered the linear regression model for various tasks, including trend estimation and testing the significance of the offset at a given change-point. In this section, we recall the two commonly used inference methods, namely, Ordinary Least Squares (OLS) and Generalized Least Squares (GLS) for this model.

The multiple linear model with k explanatory numerical variables is given, for n observations, by:

$$z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + .. + \beta_k x_{ik} + \epsilon_i \quad \text{for } i = 1, \ldots, n.$$

This model can be expressed in matrix form as:

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{X}$ is the matrix of explanatory variables with size $[n \times (k+1)]$, $\mathbf{y}$ is the $[n \times 1]$ vector of responses, $\mathbf{e}$ is the $[n \times 1]$ vector of errors and $\boldsymbol{\beta}$ the $[(k+1) \times 1]$ vector of coefficient regression, i.e. the unknown parameters:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1k} \\ 1 & x_{21} & x_{12} & . & . & . & x_{2k} \\ . & . & . & . & . & . & \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{nk} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ y_n \end{bmatrix}, \boldsymbol{e} = \begin{bmatrix} \epsilon_1 \\ . \\ . \\ \epsilon_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & . & \beta_k \end{bmatrix} \text{ respectively.}$$

The least-squares method is the most popular inference method for linear regression. It consists in finding the

value of $\boldsymbol{\beta}$ that minimizes the sum of squared errors given by

$$S(\boldsymbol{\beta}) = (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{z} - \mathbf{X}\boldsymbol{\beta}).$$

The solution, denoted by $\hat{\boldsymbol{\beta}}$, is obtained by setting the (matricial) derivative of $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ equal to 0. This leads to the equation:

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{z}.$$

When $\mathbf{X}^T\mathbf{X}$ is invertible, we obtain the following least-square (LS) estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}. \tag{2.9}$$

To study the properties of this estimator, let us substitute $\mathbf{z}$ into the previous equation. Hence:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e} \\ &= \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}.\end{aligned}$$

We can deduce some properties of the LS estimator. First, the LS estimator is unbiased as long as the noise is centered. Second, the uncertainty of the LS estimator depends on the variance of the noise (or equivalently of the variance of the response):

$$var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T var(\mathbf{z})\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}. \tag{2.10}$$

Assuming that $\mathbf{e} \sim \mathcal{N}(0, \sigma_0^2 I_n)$, where I_n is an identity matrix with size n, or at least that the errors are uncorrelated and equally variable, leads to the LS estimator with a reduced variance

$$var(\hat{\boldsymbol{\beta}}) = \sigma_0^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

This estimator is known as the Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\beta}$ (Gauss-Markov theorem). BLUE implies that the variance of the estimator is the smallest compared to other unbiased linear estimators. This solution is the familiar Ordinary Least Squares (OLS), often denoted $\hat{\boldsymbol{\beta}}^{\text{OLS}}$.

However, the OLS is no longer the BLUE when one of the two assumptions (independence and homoscedasticity) is violated, i.e. $var(\mathbf{e}) = var(\mathbf{z}) = \Sigma_0 \neq \sigma_0^2 I_n$ where Σ_0 is a positive semidefinite matrix. Indeed, in this case, we get

$$E(\hat{\boldsymbol{\beta}}^{\text{OLS}}) = \boldsymbol{\beta},$$

the OLS is still unbiased, but according to equation (2.10), we get

$$var(\hat{\boldsymbol{\beta}}^{\text{OLS}}) = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\Sigma_0\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1} \neq \sigma_0^2(\mathbf{X}^T\mathbf{X})^{-1}. \tag{2.11}$$

A serious consequence is that statistical inferences (as tests) based on the standard OLS estimation results

become invalid.

The adapted framework in this case is the General Least Squares (GLS) method which aims to minimize the following sum of squares:

$$(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T \Sigma_0^{-1} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta}).$$

The solution of this minimization is called the GLS estimator given by

$$\hat{\boldsymbol{\beta}}^{\text{GLS}} = (\mathbf{X}^T \Sigma_0^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma_0^{-1} \mathbf{z}, \tag{2.12}$$

with variance

$$var(\hat{\boldsymbol{\beta}}^{\text{GLS}}) = (\mathbf{X}^T \Sigma_0^{-1} \mathbf{X})^{-1}.$$

Note that the GLS method is equivalent to the OLS method applied on the transformed model:

$$\mathbf{G}\,\mathbf{z} = \mathbf{G}\mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{e},$$

with $\boldsymbol{G} = \Sigma_0^{-1/2}$. The GLS estimator is the BLUE of $\boldsymbol{\beta}$.

In practice, Σ_0 is unknown. Depending of the study objective, various method can be used. Two objectives can be considered: (i) when we are interesting in the estimator itself, the popular method is the feasible generalized least-squares (FGLS) estimator which consists in substituting a good estimator of Σ_0 in the (2.12) and (ii) when we are rather interesting in statistical inferences (as tests), the OLS estimator can be considered with a correction of its variance. More precisely, we can estimate the term $(\mathbf{X}^T \Sigma_0 \mathbf{X})$ in the variance (see equation (2.11)) using a robust covariance (HAC) estimator.

2.2.3 Linear stochastic models

2.2.3.1 Definition

In modeling time series, specific stationary stochastic processes that are valuable include the autoregressive (AR), moving average (MA), and mixed autoregressive-moving average (ARMA) processes (Box et al., 2016; Shumway and Stoffer, 2017). A generalization of the latter to autoregressive integrated moving-average (ARIMA) models provides also a useful class of nonstationary processes. While nonstationary processes are frequent in many application fields, in this work we are mainly concerned with stationary processes.

The autoregressive model of order p, or more succinctly, the AR(p) process, is defined by the following equation Box et al. (2016, p. 52):

$$z_t = \phi_1 z_{t-1} + \ldots + \phi_p z_{t-p} + a_t, \tag{2.13}$$

where z_t, assumed to be stationary and of zero mean, denotes the dependent variable at time t, z_{t-1} to z_{t-p} represent its previous values up to lag p, and ϕ_1 to ϕ_p are the set of model parameter. The term a_t represents an independent random sequence, or white noise, also known as the innovation. The latter follows a normal distribution with mean $\mu_a = 0$ and variance σ_a^2, i.e. $a_t \sim \mathcal{N}(0, \sigma_a^2)$. To account for a non-zero mean, μ, in z_t, a constant can be added to the right-hand side of equation (2.13), or μ can be subtracted from z_t and all its lagged

terms in the equation.

The moving average process of order q, or MA(q), is defined by the following equation Shumway and Stoffer (2011, p. 90):

$$z_t = a_t + \theta_1 a_{t-1} + ... + \theta_q a_{t-q}, \tag{2.14}$$

where a_{t-1} to a_{t-q} represent the lagged noise terms up to lag q, and θ_1 to θ_q correspond to the model coefficients associated with these lagged error terms. Note that in equation (2.14), we adopted a positive sign in the sum of the right-hand side terms, following the notation of Shumway and Stoffer (2011) and the convention in R, which is opposite to the one used in Box et al. (2016).

Finally, the definition of the ARMA(p, q) process, as formulated by Shumway and Stoffer (2011, p. 92)), is:

$$z_t = \phi_1 z_{t-1} + ... + \phi_p z_{t-p} + a_t + \theta_1 a_{t-1} + \theta_q a_{t-q}. \tag{2.15}$$

This expression represents a combination of the autoregressive model of order p and the moving average model of order q.

The AR(p) model implies that observations at time t are influenced by p previous observations with weights $\phi_1, \phi_2, ..., \phi_p$, in addition to an independent white noise term at time t. On the other hand, the MA(q) model suggests that observations at time t are affected by q previous noise terms with weights $\theta_1, \theta_2, ..., \theta_q$.

To express the time series lags in a concise mathematical form, the backward shift operator B is utilized, which is defined as follows:

$$Bz_t = z_{t-1}.$$

The AR(p) model can thus be reformulated as:

$$(1 - \phi_1 B - ... - \phi_p B^p) z_t = a_t.$$

Introducing the autoregressive polynomial $\phi(B) = (1 - \phi_1 B - ... - \phi_p B^p)$, we have:

$$\phi(B) z_t = a_t.$$

Similarly, the MA(q) model can be rewritten as:

$$z_t = \theta(B) a_t.$$

where the moving average polynomial is denoted as $\theta(B) = (1+\theta_1 B+...+\theta_q B^q)$. Finally, the mixed ARMA(p, q) model is represented as:

$$\phi(B) z_t = \theta(B) a_t.$$

2.2.3.2 Important features

A central feature in the development of time series models is the assumption of stationarity. Stationarity implies that the mean, $E[z_t]$, variance, $var[z_t] = E[(z_t - \mu)^2]$, and autocovariance between time t and $t-h$, $cov[z_t, z_{t-h}] = E[(z_t - \mu)(z_{t-h} - \mu)]$, are the same for all times t Shumway and Stoffer (2011, p. 23). For a stationary series, we have thus:

$$E[z_t] = \mu, \quad var[z_t] = \sigma_z^2, \quad \text{and } cov[z_t, z_{t-h}] = \gamma(h).$$

Note that the autocoviance depends only on the time lag h.

While the MA(q) model is always stationary, the AR(p) model does not always exhibit stationarity. It is shown in Box et al. (2016, pp. 54–55) that the AR(p) model achieves stationarity only when the roots of the autoregressive polynomial lie outside the unit circle.

Another crucial property for a linear process is invertibility. It relates to the capability of representing the error term a_t as a function of z_t, making it algebraically equivalent to a converging infinite order AR model. The AR(p) model is generally invertible, while the MA(q) model is not always so. The invertibility of the MA(q) model is achieved on the condition that $\theta(B)^{-1} \neq 0$, which implies that the roots of the moving average polynomial lie outside the unit circle (Box et al., 2016, pp. 68–69). For an ARMA(p, q) process to be stationary and invertible, both conditions must be satisfied simultaneously.

To illustrate these concepts, the following sub-section will describe the properties of three fundamental models, the AR(1), MA(1), and ARMA(1,1), that will be extensively used in this work. For simplicity, we will henceforth denote ϕ_1 and θ_1 as ϕ and θ, respectively, since the models are of first order.

2.2.3.3 Fundamental models

AR(1) model

The zero-mean first-order autoregressive model is expressed as:

$$z_t = \phi z_{t-1} + a_t,$$

where the stationarity condition restricts $|\phi| < 1$ and where $\{a_t\}_t$ i.i.d. $\sim \mathcal{N}(0, \sigma_a^2)$. When $|\phi| = 1$, it transforms into the (non-stationary) random walk model.

We can easily verify that the mean is zero by writing:

$$E[z_t] = E[\phi z_{t-1}] + E[a_t] = \phi E[z_{t-1}] = 0,$$

given that $E[a_t] = 0$ and that, due to stationarity, the mean of the process must be constant, hence $E[z_t] = E[z_{t-1}]$.

The variance can be expressed as:

$$var[z_t] = E[z_t^2] = \frac{\sigma_a^2}{1-\phi^2},$$

given that due to stationarity, the variance must be constant, i.e. $var[z_t] = \phi^2 var[z_{t-1}] + \sigma_a^2 = var[z_{t-1}]$.

The autocovariance function (ACVF) at lag h writes (Shumway and Stoffer, 2011, p. 86):

$$\gamma(h) = cov[z_t, z_{t-h}] = \frac{\sigma_a^2 \phi^h}{1-\phi^2}, h \geq 0,$$

with $\gamma(h) = \gamma(-h)$, and the Autocorrelation function (ACF) at lag h is:

$$\rho(h) = corr[z_t, z_{t-h}] = \frac{\gamma(h)}{\gamma(0)} = \phi^h, h \geq 0,$$

with the property $\rho(h) = \rho(-h)$.

The ACF tails off exponentially and rapidly approaches 0 due to the condition $|\phi| < 1$. When $\phi = 0$, the process becomes white noise, whose ACF is $\rho(h) = 0$ for all lags except for lag $h = 0$ where its value is $\rho(0) = 1$.

An important property of linear processes is causality, the property by which z_t depends only on past values, z_{t-1}, z_{t-2}, a_t, a_{t-1}... Time series issued from the physical measurements we are concerned with in this study are always causal.

In the case of the AR(1) process, causality is equivalent to, or follows from, stationarity. This can be seen by arranging the recursive equation (2.13) as a function of a_t, a_{t-1}..., or, equivalently, by inverting the polynomial equation, $(1-\phi B)z_t = a_t$, as $z_t = (1-\phi B)^{-1} a_t = \sum_{j=0}^{\infty} \phi^j a_{t-j}$. This alternative representation of the AR(1) process is known as its "infinite MA representation", which is also basically the form of any causal linear process. Stationarity of this process requires that the infinite sum is converging, which is obtained when $|\phi| < 1$.

MA(1) model

The zero-mean first-order moving average model is represented as:

$$z_t = \theta a_{t-1} + a_t,$$

where $\{a_t\}_t$ i.i.d. $\sim \mathcal{N}(0, \sigma_a^2)$. Again, we can check that the mean of this process is zero:

$$E[z_t] = \theta E[a_{t-1}] + E[a_t] = 0, \tag{2.16}$$

and the variance is

$$var[z_t] = \theta^2 var[a_{t-1}] + var[a_t] = (1+\theta^2)\sigma_a^2. \tag{2.17}$$

by virtue of the independence and stationarity of the white noise.

The ACVF at lag h writes (Shumway and Stoffer, 2011, p. 90):

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma_a^2, & h = 0, \\ \theta^2\sigma_a^2, & h = 1, \\ 0, & h > 1. \end{cases} \tag{2.18}$$

and the ACF is:

$$\rho(h) = \begin{cases} 1, & h = 0, \\ \frac{\theta}{1+\theta^2}, & h = 1, \\ 0, & h > 1. \end{cases}$$

The ACVF and ACF of the MA(1) process are equal to zero for all lags higher than 1. For the general MA(q), this holds for all lags larger than q. The MA(q) process is thus always stationary.

One caveat of the MA(q) model is the non-uniqueness of its ACVF and ACF. For the MA(1) model, for example, $\rho(h)$ is the same for θ and $1/\theta$, and $\gamma(h)$ is the same for the pair (θ, $\sigma_a^2 = 1$), and ($1/\theta$, $\sigma_a^2 = 1/\theta^2$). In practice, one chooses the model which satisfies the invertibility condition, $|\theta| < 1$ (Shumway and Stoffer, 2011, p. 92).

In the backward operator notation, the MA(1) model can be expressed as $z_t = (1 + \theta B)a_t$. When $|\theta| < 1$, the inverse of the operator $(1 + \theta B)$ is given by:

$$(1 + \theta B)^{-1} = \sum_{j=0}^{\infty} (-\theta)^j B^j.$$

This infinite sum converges for all $|B| \leq 1$. Consequently, the MA(1) model can be equivalently written as:

$$\left(\sum_{j=0}^{\infty} (-\theta)^j B^j\right) z_t = a_t.$$

This is the infinite AR representation of the MA(1) model, and it only exists when $|\theta| < 1$, i.e. when the model is invertible.

ARMA(1,1) model

The zero-mean first-order autoregressive moving average model is given by:

$$z_t = \phi z_{t-1} + \theta a_{t-1} + a_t,$$

where $\{a_t\}_t$ i.i.d. $\sim \mathcal{N}(0, \sigma_a^2)$. To be causal and unique, it necessitates both the stationarity condition of AR(1) and the invertibility condition of MA(1). Under these conditions, the ACVF at lag h writes (Shumway and Stoffer, 2017, p. 92):

$$\gamma(h) = \begin{cases} \frac{\sigma_a^2(1+2\theta\phi+\theta^2)}{1-\phi^2}, & h = 0, \\ \frac{\sigma_a^2(1+\theta\phi)(\theta+\phi)\phi^{h-1}}{1-\phi^2}, & h \geq 1. \end{cases}$$

The ACF is:

$$\rho(h) = \begin{cases} 1, & h = 0, \\ \frac{(1+\theta\phi)(\theta+\phi)}{1+2\theta\phi+\theta^2}\phi^{h-1}, & h \geq 1. \end{cases}$$

The general pattern of the ACF of the ARMA(1,1) model closely resembles that of the AR(1) model, both exhibiting exponential decay with h. Because of this similarity, relying solely on the ACF proves inadequate for distinguishing between the two models. This consideration will lead to the use of the Partial autocorrelation function (PACF).

Due to the properties of the AR(1) and MA(1), the ARMA(1,1) process can be represented either as an infinite MA process or as an infinite AR process, and these properties can actually be generalized to the ARMA(p, q) model as well.

A prevalent issue of the ARMA(p, q) models is the parameter redundancy, or over-parameterization. In the case of the ARMA(1,1) model, this can be evidenced by reformulating in the backward operator form:

$$(1 - \phi B)z_t = (1 + \theta B)a_t.$$

It is obvious that in the scenario where $\phi = -\theta$, $z_t = a_t$ which is actually a white noise process. As this example points out, we might fit an ARMA(1, 1) model to white noise data and find that the parameter estimates are significant. If we were unaware of parameter redundancy, we might claim the data are correlated when in fact they are not (Shumway and Stoffer, 2017, p. 84). This situation highlights a challenge involving the potential confusion between the white noise model and the ARMA(1,1) model that may arise during the process of model identification. This issue can be solved, in the general ARMA(p, q) model, by reducing the common factors in the $\phi(B)$ and $\theta(B)$ polynomials (Shumway and Stoffer, 2017, p. 85).

2.2.3.4 Note on the PACF

The Partial Autocorrelation Function (PACF) is a tool that helps to identify the order of an AR(p) model and, in combination with the ACF, helps to distinguish between AR, MA, and ARMA models. The idea is to construct a function that cuts off after lag p for an AR(p) model, like ACF cuts off after lag q for a MA(q) model. The principle is the following (Shumway and Stoffer, 2017, p. 96):if X, Y, and Z are random variables, then the partial correlation between X and Y given Z is obtained by regressing X on Z to obtain $\hat{X}$, regressing Y on Z to obtain $\hat{Y}$, and then calculating $\rho_{XY|Z} = corr(X - \hat{X}, Y - \hat{Y})$. The PACF of a stationary process, z_t, denoted ϕ_{hh}, for $h = 1, 2, ...$, is (Shumway and Stoffer, 2017, p. 97):

$$\begin{cases} \phi_{11} = corr(z_{t+1}, z_t) = \rho(1); \\ \phi_{hh} = corr(z_{t+h} - \hat{z}_{t+h}, z_t - \hat{z}_t), \quad h \geq 2, \end{cases}$$

where $\hat{z}_{t+h}$ is the regression of z_{t+h} on $z_{t+h-1}, z_{t+h-2}, ..., z_{t+1}$.

In the case of a stationary AR(1) process of parameter ϕ, we have:

$$\begin{cases} \phi_{11} = \rho(1) = \phi, \\ \phi_{22} = corr(z_{t+2} - \hat{z}_{t+2}, z_t - \hat{z}_t) = corr(a_{t+2}, z_t - \hat{z}_t) = 0, \\ \phi_{hh} = 0, h > 2/ \end{cases}$$

This indicates that for an AR(1) model, the PACF cuts off after the first lag. The property is the same for the AR(p) models with the PACF cutting off after lag p.

For an invertible MA(1) model of parameter θ , the PACF writes (Shumway and Stoffer, 2011, p. 107):

$$\begin{cases} \phi_{11} = \frac{\theta}{1+\theta^2}, \\ \phi_{22} = -\frac{\theta^2}{1+\theta^2+\theta^4}, \\ \phi_{hh} = -\frac{(-\theta)^h(1-\theta^2)}{1-\theta^{2(h+1)}}, h > 2. \end{cases}$$

From this equation, it is clear that the PACF for the MA(1) model will not cut off, contrary to its ACF, but instead it will tail off like the ACF of an AR(1) model. This behavior can be understood by recognizing that an invertible MA process can be represented as an infinite AR process, preventing the PACF from cutting off at any specific lag. The PACF of a causal and invertible ARMA(1,1) model exhibits the tailing off like the MA(1) model.

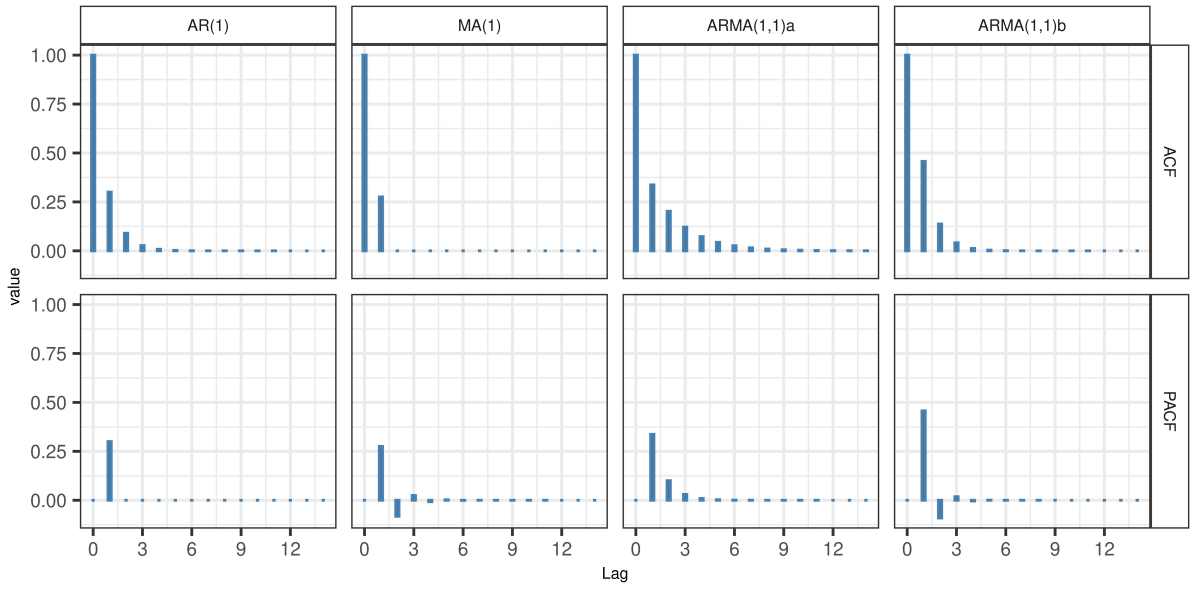


Figure 2.3 – Theoretical ACF and PACF of AR(1) model with $\phi = 0.3$, MA(1) model with $\theta = 0.3$, and ARMA(1,1) model with (a) $\phi = 0.6$ and $\theta = -0.3$, and (b) $\phi = 0.3$ and $\theta = 0.2$.

Figure 2.3 presents the ACF and PACF for the three model types with typical ϕ and θ values encountered in the real data that are discussed in this work (see Chapter 4). The three model types can be clearly distinguished. For the AR(1) model, the ACF tails off, while the PACF cuts off after lag $p = 1$. For the MA(1) model, the ACF cuts off after lag $q = 1$, while the PACF tails off. And for the two ARMA(1,1) models, both the ACF and PACF

tail off. However, depending on the sign of $\phi + \theta$, the PACF oscillates around zero or not. All these general behaviours are also valid for higher order models, AR(p), MA(q), and ARMA(p, q).

2.2.3.5 Note on spectral properties of stationary models

Time series often exhibit periodic variations which suggest that they can be approximated by a limited-order Fourier series. The concept can be generalized to non-periodic time series and linear stochastic processes by introducing a continuous distribution of frequencies (instead of harmonics considered in the simple Fourier Series). A fundamental tool in spectral analysis of stationary processes is the spectral density, also called power spectral density (PSD), $f(\omega)$, which is defined as the Fourier Transform of the ACVF (Shumway and Stoffer, 2017, p. 173):

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h}, -1/2 \leq \omega \leq 1/2, \tag{2.19}$$

where ω is the continuous frequency variable (in cycles per point). Because z_t is stationary, the inverse transform of $f(\omega)$ exists, and yields the ACVF:

$$\gamma(h) = \int_{-1/2}^{1/2} f(\omega) e^{2\pi i \omega h} d\omega, h = 0, \pm 1, \pm 2, ...$$

i.e. the PSD and the ACVF are Fourier Transforms pairs.

The fact that $\gamma(h)$ is non-negative ensures that $f(\omega) \geq 0$ for all ω, and it follows also from equation (2.19) that $f(\omega) = f(-\omega)$, and that $\gamma(0) = var[z_t] = \int_{-1/2}^{1/2} f(\omega) d\omega$.

The PSD of some special cases of interest are given below (Box et al., 2016): for $-1/2 \leq \omega \leq 1/2$,

— White Noise (WN): $f(\omega) = \sigma_a^2$,

— AR(1): $f(\omega) = \frac{\sigma_a^2}{1+\phi^2-2\phi cos(2\pi\omega)}$,

— MA(1): $f(\omega) = \sigma_a^2 \times [1 + \theta^2 + 2\theta cos(2\pi\omega)]$,

— ARMA(1,1): $f(\omega) = \sigma_a^2 \times \frac{(1+\theta^2+2\theta cos(2\pi\omega)}{1+\phi^2-2\phi cos(2\pi\omega)}$.

Figure 2.4 displays the PSDs for the same four models previously depicted in Figure 2.3. While AR(1) and MA(1) share similarities such as higher power in the lower frequencies and finite power at the limits $\omega = 0$ and $\omega = 1/2$. Note that the negative frequency range is usually not plotted due to the symmetry of the PSD ($f(\omega) = f(-\omega)$). The main difference between the AR(1) and MA(1) plots is that the PSD of the MA(1) is flatter at low frequencies and the PSD of the AR(1) decreases faster in the medium frequency range. Comparatively,the PSD of the first ARMA(1,1) model starts higher than the AR(1) and MA(1) and decreases faster the former two and than the second ARMA model. These distinguishable features can help to identify the different models from a visual inspection, as an alternative or complementary way to the ACF and PACF discussed above.

In practice, the difficulty is that real data do not necessarily follow a simple stochastic model and the usual ACF, PACF, and PSD estimators are subject to errors, making the identification more difficult.

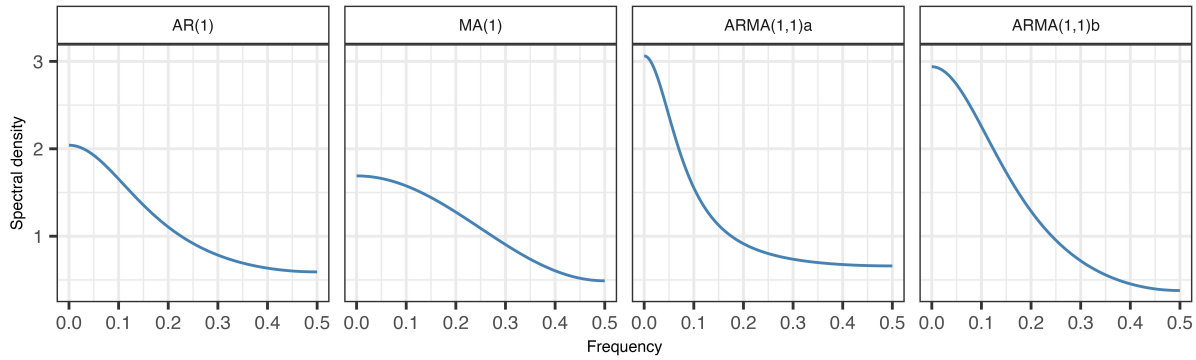


Figure 2.4 – Theoretical spectral density of AR(1) model with $\phi = 0.3$, MA(1) model with $\theta = 0.3$, and ARMA(1,1) model with (a) $\phi = 0.6$ and $\theta = -0.3$, and (b) $\phi = 0.3$ and $\theta = 0.2$. In all cases, the variance of the innovation is $\sigma_a^2 = 1$.

2.2.3.6 Combination of models

In practice, we may often consider that an observed time series is the sum of two or more independent series, e.g. a physical signal following an AR(1) process summed with an additive white noise (WN). The properties of the sum of ARMA-type models was first studied by Granger and Morris (1976). They showed that, if $X_t \sim \text{ARMA}(p, m)$, $Y_t \sim \text{ARMA}(q, n)$, and $Z_t = X_t + Y_t$, then $Z_t \sim \text{ARMA}(x, y)$, where $x \leq p + q$ and $y \leq max(p + n, q + m)$. The result can be generalized to the sum of N ARMA models:

$$\sum_{j=1}^{N} ARMA(p_j, m_j) = ARMA(x, y),$$

where

$$x \leq \sum_{j=1}^{N} p_j,$$

and

$$y \leq max(x - p_j + m_j, j = 1, ..., N).$$

Some specific combinations are of particular interest, namely:

1. AR(1) + WN = ARMA(1,1),
2. AR(1) + AR(1) = ARMA(2,1),
3. MA(1) + WN = MA(1),
4. MA(1) + MA(1) = MA(1),
5. ARMA(1,1) + WN = ARMA(1,1),
6. AR(1) + MA(1) = ARMA(1,2).

Hence, it is quite likely that real data resulting from a combination of processes give rise to ARMA models.

Granger and Morris (1976) also address the question whether a given specific ARMA(p, q) model could have arisen from simpler models. Of interest to us is, especially, the case whether an ARMA(1,1) results from the

combination of an AR(1) with WN. They show that the answer is positive when a realizability condition is met with the ARMA(1,1) model parameters, ϕ and θ, namely, if:

$$\frac{1}{1-\phi^2} > \frac{\rho_1}{\phi} \geq 0,$$

where $\rho_1 = \frac{\theta}{1+\theta^2}$ is the ACF at lag $h = 1$ of the MA(1) part of the ARMA(1,1) process, then the ARMA(1,1) = AR(1) + WN.

2.2.3.7 Model identification

The model identification for general ARMA(p, q) models aims at determinating the orders p and q. The widely used Box-Jenkins method relies extensively on a visual analysis of ACF and PACF plots (Box et al., 2016, pp. 180–185). PSD plots are sometimes used as well, although these are more commonly used for building transfer function and multivariate models (Box et al., 2016, Chap. 12).
In recent years, endeavors have been made to automate identification processes through estimation-based methods (Hyndman and Khandakar, 2008; Koreisha and Pukkila, 1995). The underlying concept involves fitting a range of potential ARMA models (different values of couple (p, q)) and then selecting the optimal one using a penalized version of the fitted criterion. The most common penalty criteria are the Akaike information criterion (AIC), bias corrected AIC (AICc), and Bayesian information criterion (BIC). The AIC is defined for a model m as:

$$AIC(m) = -2\ \ln(\hat{L}(m)) + 2\ D(m),$$

and the BIC as:

$$BIC(m) = -2\ \ln(\hat{L}(m)) + D(m)\ \ln(n),$$

where $\hat{L}(m)$ denotes the estimated likelihood of model m, $D(m)$ is its number of parameters to be estimated, and n is the sample size. Since the BIC penalty is higher than the AIC one, BIC tends to select simplest or parsimonious models compared to AIC (i.e. small values of p and q). Note that for ARMA framework, AIC is known to overestimate the orders (Shibata, 1976).

In our study, we consider the BIC criterion and use the "auto.arima" function with $d = 0$ of the R package "forecast" (Hyndman et al., 2018).

2.2.3.8 Parameter estimation

To estimate the coefficient parameters in ARMA models, the well known methods proposed in the literature are the maximum likelihood, the conditional or unconditional ones (or equivalently the conditional or unconditional Sum of Squares), and for the specific case of AR models, the method of moments using the Yule-Walker equations. The exact or unconditional likelihood leads to a complex optimization problem since it is non-linear according to the parameters and thus requires the use of iterative algorithms. The conditional likelihood (conditionally to the initial observations) method is a way to simplify the problem. However, even if for very large samples, the both likelihood estimators are equivalent with the same asymptotic distribution (Hamilton 1994, p. 126), the conditional likelihood can result in biased estimates for relatively short series (Box et al., 2016, p. 526). Because of this, the use of the unconditional likelihood function is typically recommended for models

with moving average terms. For a complete presentation and discussion of these inference methods, see Box et al. (2016, Chap. 7) or Brockwell et al. (1991, Chap. 8). In our study, we consider the unconditional inference likelihood approach and use the "arima" function in R.

The large sample variance for the maximum likelihood estimates of an AR(1), MA(1) and ARMA(1,1) models are (Box et al., 2016, pp. 233–238):

AR(1) model:

$$var[\hat{\phi}] \simeq \frac{1-\phi^2}{n}, \tag{2.20}$$

MA(1) model:

$$var[\hat{\theta}] \simeq \frac{1-\theta^2}{n}, \tag{2.21}$$

ARMA(1,1) model:

$$var[\hat{\phi}] \simeq \frac{(1+\phi\theta)^2(1-\phi^2)}{n(\phi+\theta)^2}, \tag{2.22}$$

$$var[\hat{\theta}] \simeq \frac{(1+\phi\theta)^2(1-\theta^2)}{n(\phi+\theta)^2}. \tag{2.23}$$

These equations underscore the relationship between the variance of coefficient estimates, the length of series, n, and the values of the coefficient, ϕ and θ. As one could guess, the variances are inversely proportional to n and are decreasing when the parameters ϕ or θ increase.

To quantify, let's consider an AR(1) process with a coefficient value of $\phi = 0.3$ and a series of length $n = 1000$. According to equation (2.20), $var[\hat{\phi}] = (0.03)^2$, signifying a relative uncertainty in $\hat{\phi}$ of 0.03/0.3=0.1, i.e. 10%. However, if we maintain the same series length while increasing the coefficient value to $\phi = 0.6$, the variance reduces to $(0.025)^2$, which corresponds to a relative uncertainty of 4%.
Now let's consider a scenario where the model is identified as an ARMA(1,1) but the true model is AR(1), with ϕ=0.3. In this case, $var[\hat{\phi}]$, as given by equation (2.22), amounts to $var[\hat{\phi}] = (0.042)^2$, which is 1.96 times larger than the variance of the AR(1) model of $(0.03)^2$. This illustrates the over-fitting caveat and emphasizes the necessity of identifying the appropriate stochastic model rather than relying on a more general model.

2.2.4 Classification

Classification is a supervised machine learning method whose aim is to predict the membership of an individual to a given class of the population, summarized by the class number y called the label, from the information available on this individual x. Solving a classification problem amounts to construct a classifier, i.e. a function ψ such that $\psi(x)$ represents the prediction of y given x. We naturally want to build a high-performance classifier, whose prediction error rate is as low as possible. Assume that (x, y) is a realization of the random vector (X, Y) of (unknown) distribution P. The classifier makes an error when $\psi(X) \neq Y$, and the quality of the classifier ψ can be measured thus by the probability of misclassification:

$$\mathcal{R}_P(\psi) = E[\mathbb{1}_{\{\psi(X)\neq Y\}}] = P(\psi(X) \neq Y),$$

This probability is also known as the risk of ψ. The ideal or optimal classifier will be the one that minimizes the risk $\mathcal{R}_P(\psi)$ denoted $\psi^\star$. This classifier is particular and is called the Bayes classifier ($\psi^\star = \arg\max_k P(Y = k|X = x)$). However, $\psi^\star$ depends on the distribution P of (X, Y) which is unknown in practice. The objective of the supervised classification is thus to construct a learning rule or classifier from a sample $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ with the same distribution as (X, Y) and with performance close to $\psi^\star$ in terms of risk. This classifier is denoted $\widehat{\psi}(D_n)$.

There are four popular learning algorithms in the literature: the Classification and Regression Tree (CART), the Random Forest (RF), the k nearest neighbors (k-NN) and the Linear Discriminant Analysis (LDA). The first three algorithms are non-parametric, in contrast to LDA. We describe these algorithms in subsection 2.2.4.1. The performance of a learning algorithm is summarized by the risk of its resulting classifier, $\mathcal{R}_P(\widehat{\psi}(D_n)))$. However, again the distribution P being unknown, this risk can not be calculated. We can estimate it by its empirical version on the sample D_n:

$$\widehat{\mathcal{R}}_n(\psi; D_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\psi(X_i)\neq Y_i\}}, \tag{2.24}$$

This is referred to as the empirical risk, representing the number of errors the classifier ψ makes on the sample D_n. It serves as an unbiased estimator of the risk associated with ψ, i.e. $E[\widehat{\mathcal{R}}_n(\psi; D_n)] = \mathcal{R}_P(\psi)$.

The quality of the classifier $\widehat{\psi}(D_n)$ is thus given by $\widehat{\mathcal{R}}_n(\widehat{\psi}(D_n); D_n)$. However, this measure of quality is biased because it is associated to an optimistic estimation of the risk $\mathcal{R}_P(\widehat{\psi}(D_n))$. As clearly indicated by these notations, the sample D_n is used twice: once to construct the classifier and once for its evaluation. To control this optimism or reduce the bias, several strategies have been proposed, including cross-validation and bootstrapping. Their objective, in particular, is to estimate the risk without bias or with reduced bias by avoiding using the same data to construct the classifier and to study its performance. Some versions of these resampling methods are presented and also discussed according to the properties of the proposed risk estimators (the bias and the variance, $E[\widehat{\mathcal{R}}]$ and $V[\widehat{\mathcal{R}}]$, respectively, of an estimator $\widehat{\mathcal{R}}$) in subsection 2.2.4.2.

Finally in subsection 2.2.4.3, we discuss the imbalance class problems for such classification purpose.

Note that this classification framework is used in the attribution method developed in Chapter 4 of this thesis.

2.2.4.1 The popular learning algorithms

We assume that the vector X is composed of p variables $X^1, X^2, \ldots, X^p$ and recall that Y is the label, i.e. the target variable.

Classification and Regression Trees (CART)

CART is a decision tree algorithm designed to create a tree-like structure by recursively partitioning the data. The root node of the tree contains all the data, each internal node corresponds to a split of the data into two

subsets and the terminal node, called leaves, is associated to a class label. The paths from root to leaf represent classification rules. An example of a classification tree is given in Figure 2.5 with comments in the caption.

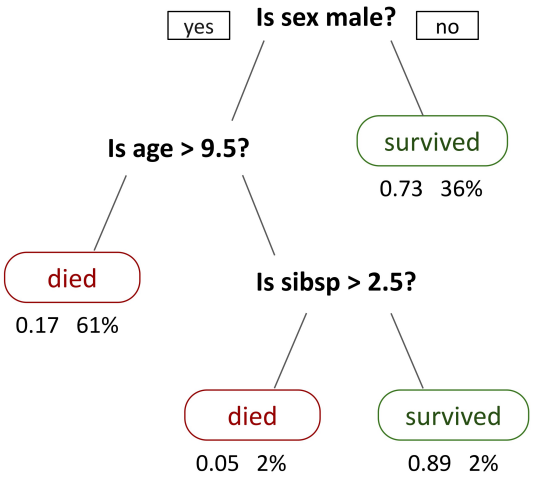


Figure 2.5 – A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The number in the leaves give the probability of survival and the percentage of observations. If the passenger is a male at most 9.5 years old with strictly fewer than 3 siblings, he will be classified as a survivor (Wikipedia contributors, 2023).

The construction of the tree therefore consists in providing a sequence of nodes where each node is defined by a question, i.e. the joint choice of a variable X^j among the p variables and a threshold d of the quantitative variable X^j:

$$\{X^j \leq d\} \cup \{X^j > d\}.$$

Splitting the data according to this decision rule means that the observations with a value of the jth variable smaller than d go to the left child node, and all those with a value larger than d go to the right child node (the two branches of the node). The objective is to make the best splitting, i.e. to search the best combination variable-threshold making the two resulting subsamples as pure/homogeneous as possible according to a criterion. The CART algorithm thus requires:

- ⋆ a "purity" or "homogeneity" criterion. Common metrics include the Gini index, misclassification error, and entropy. In this study, we employ the Gini index given by $Gini = 1 - \sum_{k=1}^{K} p_k^2$ where K is the number of classes and p_k corresponds to the probability of an individual being classified to class k.
- ⋆ a stopping rule since this process only halts when further splitting of a node is no longer feasible such as when each terminal node (leaf) contains only one observation.
- ⋆ a leaf assignment. Lastly, each terminal node is assigned to a class label.

Choosing a stopping rule a priori is a challenging task. In practice, we construct a complete tree called the maximal tree and denoted $T_{\max}$, that is to say a tree large enough to classify without error all the data in the training sample (leading inevitably to overfitting). Then we search for the best pruned subtree of $T_{\max}$, i.e. from which we remove the branches that do not provide significant predictive power. An exhaustive search (considering all the subtrees) is not possible from an algorithmic point of view since their number is exponential. To get around this problem, Breiman et al. (1984) proposed to construct a sequence of nested

subtrees $T_1 \preceq T_2 \preceq \ldots \preceq T_{\max}$ and to select the one that minimizes the following penalized misclassification error:

$$\widehat{\mathcal{R}}_n(T_i) + \alpha|T_i|,$$

where $|T_i|$ represents the size of the tree T_i and $\widehat{\mathcal{R}}_n(T_i)$ its empirical misclassification.

In conclusion, CART is a non-parametric method that is simple to understand, interpret, and visualize. The drawback is that it is particularly unstable, i.e. very sensitive to fluctuations in the sample (small variations in the data might result in a completely different decision tree).

Random Forest (RF)

In order to overcome the issue of instability in CART, Breiman (1996) introduced the Random Forest (RF) algorithm. The idea of this algorithm is to consider multiple decision trees to improve the predictive power but at the cost of a loss of interpretability. According to the Strong Law of Large Numbers, the error rate of the ensemble of trees converges as the number of trees increases (Breiman, 2001). This convergence effectively mitigates the overfitting problem commonly associated with single decision trees, thereby enhancing both the model's accuracy and efficiency.

The underlying principle of RF can be succinctly explained by the Bagging (Bootstrap Aggregating) technique, which is a method for generating multiple versions of a predictor on bootstrap samples and using these to get an aggregated predictor (Breiman, 1996) (as illustrated in Figure 2.6). Further details on bootstrap sampling are provided in subsection 2.2.4.2, and the typical approach to result aggregation is by majority vote.

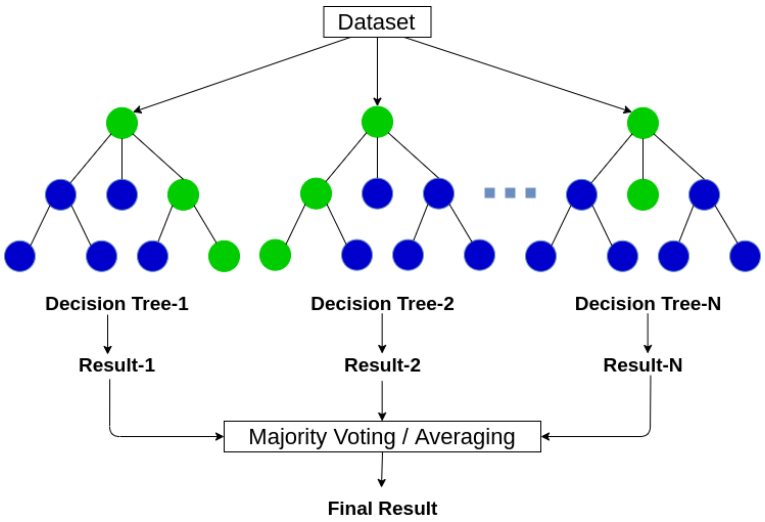


Figure 2.6 – Random forest illustration (TIBCO, 2023).

In Random Forest, the decision rule (question) at each node is determined on the basis of a random subsample of variables m out of p variables. This additional layer of randomness serves to further reduce the correlation between the decision trees, reducing the variance of the aggregated predictions and thereby enhancing the bagging. To quantify the risk of the ensemble, Breiman (2001) introduced the concept of out-of-bag (OOB)

estimation, which is elaborated upon in subsection 2.2.4.2.

Random Forest offers several advantages, including enhanced predictive accuracy, and flexibility in ensemble composition, as it can incorporate various types of classifiers. However, it is essential to acknowledge that Random Forest can be computationally expensive, especially as the number of trees in the forest grows. Additionally, the aggregation of results from multiple trees can make the interpretation of the model more challenging compared to a single decision tree.

k Nearest Neighbors (k-NN)

Another powerful non-parametric classification algorithm is the k-NN (Fix and Hodges, 1989; Cover and Hart, 1967). The intuitive and simple idea behind k-NN is to assume that very close observations probably belong to the same class. Figure 2.7 illustrates the k-NN procedure. Initially, we have data in classes A and B and need to classify an unknown observation, represented by the purple circle with a question mark, into one of these classes. The k-NN procedure involves the two following steps for classifying a new observation x:

1. Identify the k neighbors based on their distance to x (the distance metric can be the Euclidean distance, the Manhattan distance, or any other suitable measure). In the right plot of Figure 2.7, $k = 3$ neighbors are chosen.
2. Determine the predicted class of x by the majority class among the k identified neighbors. For this example, x is classified into class B.

Figure 2.7 – k-NN illustration (IBM, 2023).

This algorithm is straightforward and does not require a separate training phase, making it relatively easy to implement. However, this method becomes computationally expensive when applied to large datasets. Computing the distance between observations becomes computationally complicated as the number of variables increases. Additionally, it tends to exhibit poorer performance on imbalanced datasets, where certain classes have substantially more observations than others. In such cases, k-NN may favor the majority class during prediction, leading to biased outcomes. The choice of k is of course crucial for the estimation quality: a small k may result in overfitting, while a large k may lead to underfitting. This number is classically chosen via cross-validation.

Linear Discriminant Analysis (LDA)

Introduced in 1936 by Fisher (1936), Linear Discriminant Analysis (LDA) was initially designed for two-class problems. The objective of the LDA is to estimate, within a parametric framework, the Bayes classifier, $\psi^\star$. More precisely, the rule of the Bayes classifier is to classify an observation x in the most probable group, i.e. the group that maximizes the posterior probability of group membership given by

$$P(Y = k|X = x).$$

This conditional probability can be decomposed using the Bayes formula as follows:

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_l \pi_l f_l(x)},$$

where the $\pi_k = P(Y = k)$ are the prior probabilities of group membership and the f_k are the density distributions of $X|Y = k$.

The LDA assumes that the distribution of X in each class is gaussian

$$X|Y = k \ \sim \ \mathcal{N}(\mu_k, \Sigma),$$

where $\mu_k \in \mathbb{R}^p$ and Σ is a definite positive matrix with size $p \times p$. It is straightforward to show that for two classes k and l,

$$B(x) = \log\left(\frac{P(Y = k|X = x)}{P(Y = l|X = x)}\right) = \log\left(\frac{f_k(x)\pi_k}{f_l(x)\pi_l}\right) = \boldsymbol{b^T} x + b_0,$$

with $\boldsymbol{b} = \boldsymbol{\Sigma^{-1}}(\boldsymbol{\mu_k} - \boldsymbol{\mu_l})$ and $b_0 = -\frac{1}{2}(\boldsymbol{\mu_k^T}\Sigma^{-1}\boldsymbol{\mu_k} - \boldsymbol{\mu_l^T}\Sigma^{-1}\boldsymbol{\mu_l}) + \log(\pi_k/\pi_l)$. Between these two classes, an observation will be classified to class k if $B(\boldsymbol{x}) > 0$. The decision boundary between classes k and l is thus a linear function of x.
The prior probabilities π_k and the distribution parameters μ_k and Σ are estimated using the classical maximum likelihood method.

LDA offers multiple advantages such as computational efficiency, simplicity, and can work well even if the number of variables is large. Moreover, it can take into account multicollinearity (correlation between variables) in the data. The limitations of LDA mainly relate to the normality assumptions and the greater the deviation from this hypothesis, the less it will be possible to guarantee the performance of the chosen classifier. One can choose to consider that the covariance matrice is class-specific. In this case, the decision boundary is rather a quadratic function of x, called the QDA. The QDA is more flexible than the LDA but at the price of a high cost of the estimation.

Parameters of the algorithms

All these algorithms involve parameters that the user must chosen (the penalty constant α in CART, the number of considered variables at each node m for RF or the numbr of neighbors for kNN). Generally, these parameters are calibrated via cross-validation strategy applied on a predefined grid of values for the parameter. Then the

optimal value is the one that minimizes the estimated risk. In this thesis, we use the $K = 10$-fold cross-validation and the R package caret.

2.2.4.2 Resampling methods

As outlined in the introduction, the two most common used resampling methods to evaluate the performance of a classifier or model are cross-validation and bootstrap.

Cross validation

The general principle of cross-validation (CV) involves dividing the sample D_n into two sub-samples: a training sample D_n^L used to train a classifier and a testing sample D_n^T (the remaining data, $D_n^T = (D_n^L)^c$ with size n_T) used to measure the performance of this classifier. Due to the independence between D_n^L and D_n^T, we obtain a good assessment of the risk of $\widehat{\psi}(D_n)$ (avoiding the over-optimism of the empirical risk) given by

$$\hat{\mathcal{R}}^{\text{val}}(\widehat{\psi}; D_n; L) = \frac{1}{n_T} \sum_{i \in T} \mathbb{1}_{\{\widehat{\psi}(D_n^L, X_i) \neq Y_i\}}.$$

Since various strategies can divide the same sample, there exists a large number of possible validation procedures. The common ones are:

- ⋆ the "hold-out" validation or simple validation, which involves a single division of D_n. This procedure can be repeated K times, each time changing the validation sample. One limitation, however, is that this procedure does not guarantee consistent data dependencies in the test set across iterations.
- ⋆ the "K-fold" cross-validation, which consists in dividing the sample D_n into K blocks of observations of equal size, each block serving in turn as a validation set and the remaining $K - 1$ blocks composing the training sample. Note that the known leave-one-out cross-validation (LOOCV) is a particular case of the K-fold CV with $K = n$. In practice, K needs to be chosen (by the user) and this choice is crucial: a smaller K yields a higher bias of the risk estimator, but the variance is concurrently diminished. The well-known trade-off between bias and variance is detailed in Rodriguez et al. (2010).

The risk evaluated by CV is then the average of the K risks:

$$\widehat{\mathcal{R}}^{\text{CV}} = \frac{1}{K} \sum_{k=1}^{K} \hat{\mathcal{R}}^{\text{val}}(\widehat{\psi}; D_n; L(k)),$$

where $L(k)$ is the kth set of observations belonging to the training set D_n^L, i.e. the training set of the kth repetition in the repeated hold-out validation or the training set including all blocks except the kth one (on which the constructed classifier is evaluated) in the K-fold CV.

As Arlot (2018) explains in his interesting document: "Intuitively, we can already say that K-fold cross-validation has the advantage of making a “balanced” use of the data: each observation is used exactly $K - 1$ times for training and once for learning. This is by no means guaranteed with repeated hold-out validation. On the other hand, one can wonder about the drawbacks of always using together (either for training or for validation) the

observations of the same block. The repeated hold-out approach, due to its random nature, makes it possible to avoid any biases induced by this link between observations".

Bootstrap

In contrast to cross-validation procedures, which partition the original data to train and evaluate, the bootstrap method (Efron and Tibshirani, 1993) involves random sampling with replacement from the original dataset to create new datasets. These datasets, known as bootstrap samples of the same size, are used to construct classifiers, and then combined to evaluate the risk.

More precisely, the ordinary bootstrap procedure consists in repeating the following steps B times:

- ⋆ create a new dataset, denoted D_n^b, from the original one, D_n, by randomly sampling D_n with replacement,
- ⋆ construct a classifier $\widehat{\psi}^b = \widehat{\psi}(D_n^b)$,
- ⋆ calculate its risk on D_n: $\widehat{\mathcal{R}}_n(\psi^b; D_n)$ (where $\widehat{\mathcal{R}}_n$ is given by (2.24)).

Combining all the B risks, the estimation of the risk by bootstrap is therefore:

$$\widehat{\mathcal{R}}^{\text{Boot}} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\mathcal{R}}_n(\hat{\psi}; D_n).$$

This estimator is generally biased due to optimism, but it does offer enhanced precision in the estimation of the risk.

To counter this optimistic bias, Breiman (2001) proposed the out-of-bag (OOB) bootstrap estimator. During sampling, certain data points are replicated within each "bag", leading to a portion of the original data being excluded and forming the "out-of-bag" set. The classifier derived from each sample is subsequently tested on its associated "out-of-bag" set. The risk estimator becomes:

$$\widehat{\mathcal{R}}^{\text{OOB-B}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{B_i} \sum_{b \in K_i} \mathbb{1}_{\{\psi^b(X_i) \neq Y_i\}}.$$

where K_i indicates the set of bootstrap samples that do not contain the observation i and B_i is its size $B_i = |K_i|$.

Summary

Both cross-validation and bootstrap methodologies are valuable for evaluating model performance, but they have different strengths and weaknesses. If the cross-validation is a relatively simple procedure to understand and implement, it can be computationally demanding especially for large datasets, and of course it is sensitive to the choice of the folds and their number. Bootstrap is a more flexible method but it is conceptually more difficult to understand and thus less used compared to CV. Note that for equal computational effort $K = B$, the bootstrap provides an estimator with a lower variance and the cross-validation with a lower bias.
In some scenarios, a combined approach can yield more robust error estimations. For instance, the leave-one-out bootstrap method, discussed by Efron and Tibshirani (1997) exhibits reduced variance compared to the

traditional leave-one-out cross-validation (LOOCV) approach.
Similarly, the concept of "bootstrap cross-validation", initially introduced by Fu et al. (2005), involves employing cross-validation within each bootstrap sample. This method offers multiple advantages, including an unbiased estimator of cross-validation for each individual sample and the ability to effectively handle imbalanced datasets through the utilization of the bootstrap strategy.

2.2.4.3 Imbalance class problem

One important problem in classification arises when the classes are imbalanced. The imbalance is often inherent to the problem at hand, as for example for medical diagnosis of a rare disease. In such situation, the learning algorithms will have difficulty to correctly identify the minority classes (the resulting classifier being strongly biased for the minority classes).

To address this problem, a typical approach involves a resampling technique that aims to rebalance the classes. Random oversampling of the minority class or undersampling of the majority class can be employed to modify the data and create a more balanced data set. While resampling can be a simple and straightforward solution that can improve classification performance, it can also introduce potential issues. Oversampling can lead to overfitting, as duplicate data is introduced, while undersampling can cause underfitting by removing representative data. In addition to resampling techniques, He and Garcia (2009) has reviewed other methods such as Cost-Sensitive methods, Kernel-based methods, and Active Learning methods. Among these methods, Cost-Sensitive methods prove to be particularly valuable as they take into account the cost of misclassifying instances in the minority class more than those in the majority class. This approach can be more effective than sampling methods, especially when the minority class carries greater significance. However, achieving optimal performance with Cost-Sensitive methods requires careful tuning of the cost parameters.

Chapter 3

Sensitivity of Change-Point Detection and Trend Estimates to GNSS IWV Time Series Properties

3.1 Summary

This chapter is presented in the form of a published article. The objective of the study is to investigate the sensitivity of the segmentation method to data properties and assess the consequential impact of these properties on trend estimations. We undertake this investigation by comparing pairs of datasets based on four critical factors: GNSS data processing methods (IGSrepro1 vs. CODE REPRO2015), the temporal extent of the time series (17 years vs. 25 years), the auxiliary data utilized in the Integrated Water Vapor (IWV) conversion, and the sources of reference data (ERAI and ERA5). Our analysis places particular emphasis on specific data properties, including mean values, noise levels, and periodic bias, as these properties have an impact on the segmentation results, particularly in terms of their influence on the number and positioning of change-points. This investigation allows us to both comprehend the segmentation results and identify important features of GNSS that can create inhomogeneities and impact trend estimates.

Our findings indicate significant impacts on the segmentation results when altering the GNSS processing and reference reanalysis. Only $45-49\%$ of change-points are similar in these cases, compared to $71-81\%$ similarity for the other two remaining factors (temporal extent and auxiliary data). Notable changes in GNSS processing include adjustments in the a priori ZHD correction, antenna/radome calibration model, and mapping function. Improvements in CODE processing lead to noise reduction and decreased periodic bias. Similarly, transitioning from ERAI to ERA5 as a reference reduces representativeness errors, resulting in noise and periodic bias reduction, making it easier to detect smaller change-points. The validation rate of the detected change-points with respect to metadata was found to consistently falls within the range of 30% to 35% for all datasets, i.e. the impact of data properties on that metric is small. This result suggests that the main instrumental changes that impact the GNSS IWV estimates were captured by the segmentation in both datasets.

The impact on trend estimations is investigated in two aspects. Firstly, we examine the influence of the four

factors previously explored in segmentation on IWV trend estimates. We find that changing the length of the time series has the strongest impact, both on the average and dispersion of the trend estimates across the network. The change in mean is believed to reflect the intensification of the water cycle in the 2010-2020 decade compared to the previous one (see the steeper global mean IWV trend in Figure 1.3). The reduced dispersion with the longer period is mainly a result of the smaller standard error of the estimates. Secondly, we consider the impact of inhomogeneity correction on trend estimates. We find that it is particularly pronounced when only change-points validated with respect to GNSS metadata (receiver, antenna, radome changes) are corrected. This impact is evident in global mean trends, dispersion, and RMS differences with respect to ERA5. Correction using all change-points also impacts the trend estimates but to a slightly lesser extent. This underscores the importance of addressing the attribution step, which aims at retaining only the change-points due to GNSS for the correction. The development of this step is the topic of Chapter 4. Finally, the dispersion associated with different homogenized trend estimates can be used as a measure of the uncertainty in the trend estimate at a single station. It amounts to $0.1 - 0.2$ kg m^{-2} decade^{-1}, or 0.5-1% decade^{-1}, which confirms the feasibility of detecting relevant global and regional climate trends with the GNSS IWV data.

3.2 Paper No. 1

atmosphere

MDPI

Article

Sensitivity of Change-Point Detection and Trend Estimates to GNSS IWV Time Series Properties

Khanh Ninh Nguyen [1,2,*,†], Annarosa Quarello [1,2,†], Olivier Bock [1,2,†] and Emilie Lebarbier [3,†]

[1] Institut de Physique du Globe de Paris (IPGP), Centre National de la Recherche Scientifique (CNRS), Institut National de l'Information Géographique et Forestière (IGN), Université de Paris, 75005 Paris, France; annarosa.quarello@edu.unito.it (A.Q.); bock@ipgp.fr (Q.B.)
[2] Ecole Nationale des Sciences Géographiques (ENSG), Institut National de L'information Géographique et Forestière (IGN), 77455 Marne-la-Vallée, France
[3] Laboratoire Modal'X, UPL, Université Paris Nanterre, 92000 Nanterre, France; emilie.lebarbier@parisnanterre.fr
* Correspondence: knguyen@ipgp.fr
† These authors contributed equally to this work.

Abstract: This study investigates the sensitivity of the GNSSseg segmentation method to change in: GNSS data processing method, length of time series (17 to 25 years), auxiliary data used in the integrated water vapor (IWV) conversion, and reference time series used in the segmentation (ERA-Interim versus ERA5). Two GNSS data sets (IGS repro1 and CODE REPRO2015), representative of the first and second IGS reprocessing, were compared. Significant differences were found in the number and positions of detected change-points due to different a priori ZHD models, antenna/radome calibrations, and mapping functions. The more recent models used in the CODE solution have reduced noise and allow the segmentation to detect smaller offsets. Similarly, the more recent reanalysis ERA5 has reduced representativeness errors, improved quality compared to ERA-Interim, and achieves higher sensitivity of the segmentation. Only 45–50% of the detected change-points are similar between the two GNSS data sets or between the two reanalyses, compared to 70–80% when the length of the time series or the auxiliary data are changed. About 35% of the change-points are validated with respect to metadata. The uncertainty in the homogenized trends is estimated to be around 0.01–0.02 kg m^{-2} $year^{-1}$.

Keywords: segmentation; homogenization; climate; GNSS; integrated water vapor; time series; trend; reanalysis

Citation: Nguyen, K.N.; Quarello, A.; Bock, O.; Lebarbier, E. Sensitivity of Change-Point Detection and Trend Estimates to GNSS IWV Time Series Properties. *Atmosphere* **2021**, *12*, 1102. https://doi.org/10.3390/atmos12091102

Academic Editor: Peter Domonkos

Received: 30 July 2021
Accepted: 20 August 2021
Published: 26 August 2021

1. Introduction

Long records of observational data are essential to monitoring climate change and understanding the underlying climate processes [1,2]. However, long time series are often affected by inhomogeneities due to changes in instrumentation, in station location, in observation and processing methods, and/or in the measurement conditions around the station [3]. Inhomogeneities often take the form of abrupt changes, which are detrimental to estimating trends and multi-scale climate variability [4]. Various homogenization methods have been developed for the detection and correction of such change-points in the context of climate data analysis [1,2,5–9].

In this paper, we are interested in ground-based Global Navigation Satellite System (GNSS) integrated water vapor (IWV) measurements. GNSS measurements are qualified among the most accurate and continuous IWV measurements in all weather conditions but have only quite recently been considered for climate analysis [10–13]. Parracho et al. [14] was one among the first to analyze global IWV trends from more than 15 years of GNSS data and to confront them to the ECMWF reanalysis, ERA-Interim (ERAI), [15], and to the NASA/MERRA-2 reanalysis [16]. Significant differences were discovered in IWV trends

between the two reanalyses and between the reanalyses and the GNSS data. On the one hand, this study pointed to the importance of the atmospheric model, the assimilation system, but also the quality and quantity of assimilated observations in reanalyses. On the other hand, inhomogeneities were also suspected in the GNSS data at several sites. Developing a homogenized GNSS IWV time series is of prime importance to estimate regional and global IWV trends and variability but also to verify climate models and reanalyses. This study investigates in more detail the homogeneity of the GNSS IWV data set used by Parracho et al. [14], as well as a more recently reprocessed GNSS data set. It also updates the previous results from Parracho et al. [14] and Bock and Parracho [17] with the new ECMWF reanalysis, named ERA5 [18].

The main causes of inhomogeneities in GNSS IWV time series are:

- Equipment changes (antenna, radome, and receiver). Each antenna/radome pair has a particular impact on the measurements, which is taken into account at the processing level with a specific calibration model (see Section 2). However, model imperfections, multipath and on-site electromagnetic coupling with the antenna's environment, and equipment aging are responsible for small biases which can change over time. The quality of measurements also depends on the receivers. Modern receivers have more stable clocks, reduced cycle slips, and noise and are capable of observing satellites from new GNSS systems (GPS, GLONASS, etc.). Hence, changes in data quality/properties are expected, which can introduce offsets and possibly trends (e.g., when new satellites are introduced progressively). Changes in receiver settings, such as cutoff angle, are also known to produce abrupt changes in the mean IWV estimates [19].
- Changes in the environment near the receiver antenna can introduce multipath and obstructions that alter the measurements and cause inhomogeneities.
- Processing changes. The details of the data processing are known to impact the IWV estimates. The most important aspects and parameters are the tropospheric model (the mapping functions, the a priori hydrostatic model, the time-dependency), the antenna/radome calibration models, the elevation-dependent weighting, and the cutoff angle (see Section 2).

The first cause is well documented for International GNSS Service (IGS) stations and other scientific networks (ftp://igs.ign.fr/pub/igs/igscb/station/log/, accessed on 30 July 2021). Therefore, metadata can be used to check if change-points detected in the IWV time series can be explained by known equipment changes. The second cause is usually not well documented, but the analysis of the raw measurements and post-fit residuals can help to detect changes in the environment. The third cause is of a different nature as it depends on the analysis procedure and models, which are both the subject of active research in order to improve the accuracy and homogeneity of the GNSS products (see Section 2). However, not all biases and inhomogeneities can be corrected at the processing level, and further post-processing homogenization methods are needed.

Many different homogenization methods have been developed by climatologists. The heart of any homogenization method is the detection of change-points, the so-called segmentation method. Some segmentation methods use statistical tests [6,7], while others use a penalized likelihood approach [1,2,20]. The performance of both approaches are comparable, but, in general, the results depend on the data properties (nature of the background noise, presence of a periodic bias and/or a trend), the adopted model (parametric or non-parametric), and the search method (optimal or sub-optimal) [9,21].

Quarello [21] developed a segmentation method, called GNSSseg, especially devoted to detect changes in the mean of time series of IWV differences between GNSS and a reference and taking into account the presence of a periodic bias and a heterogeneous noise with a monthly variation. The method uses a penalized likelihood approach and is optimal in the sense that the estimation of the positions of the change-points is done using an efficient algorithm. The method proposes several penalty criteria, which aims to choose the number of change-points, with different sensitivities to the data properties (length of

the time series, noise distribution, etc.). The use of several criteria can help to mitigate their limitations but requires special post-processing to make the final decision, either automatic or manual. The post-processing may also include outlier detection, validation with metadata when available, and manual inspection. The automatic version of the GNSSseg algorithm was evaluated in a benchmark exercise and compared to other existing segmentation methods where it was found to be one of the most efficient in detecting change-points in synthetic time series mimicking the GNSS minus ERAI IWV differences at the moderate complexity [22].

The general objective of this paper is to evaluate the sensitivity of segmentation results with the GNSSseg method (recently improved in terms of computational time and so-called GNSSfast method) and the subsequent trend estimates to various qualitative and quantitative properties of both GNSS and reference data. The study considers the particular cases of two different GNSS data sets (IGS repro1 and CODE REPRO2015) combined with two different reanalysis data sets, ERA-Interim [15] and ERA5 [18], which serve as references to compute to IWV differences used in the segmentation. IGS repro1 and CODE REPRO2015 are representative of the 1st and 2nd generation of IGS reprocessing products, and, as such, they are expected to be of different quality. They also cover different time periods. ERAI and ERA5 are the 4th and 5th generation reanalyses produced by ECMWF [18] and are also of different quality and spatial resolution.

The paper is organized as follows. In Section 2, we describe the characteristics of the two GNSS IWV data sets and discuss which factors in the data processing control the accuracy of the daily IWV estimates and their homogeneity in the long term. We also present the global homogenization and the trend estimation methods. In Section 3.1, we study the impact of data properties on the segmentation results. The following questions are specifically investigated: (1) What is the impact of the different data processing between IGS repro1 and CODE REPRO2015 on the segmentation results? (2) what is the impact of the time length on the segmentation results? (3) What is the impact of the reference data source (ERAI versus ERA5)? (4) What is the impact of the auxiliary data source used in the ZTD to IWV conversion (ERAI versus ERA5)? The segmentation results will be compared on the basis of several statistics: the number of detections before and after outlier screening, the number of outliers, and the number of validations/attributions after screening using metadata and nearby stations. In Section 3.2, the corrected time series are used to estimate linear trends, and the same questions as above are investigated for the trends. Finally, Section 4 concludes the paper and discusses the future work.

2. Materials and Methods

2.1. GNSS IWV Data

Before describing the characteristics of the two GNSS IWV data sets used in this study, we deem it is necessary to discuss which factors in the data production control the accuracy of the daily IWV estimates and their homogeneity in the long term. The raw GNSS measurements consist of code and carrier phase signals transmitted by GNSS satellites (GPS, GLONASS, etc.), which are measured by ground-based stations composed of an antenna and a multi-channel receiver [23]. The measurements collected from a global tracking network are analyzed in routine by several International GNSS Service (IGS) analysis centers who compute accurate satellite orbits and clock parameters (https://www.igs.org/acc/, accessed on 30 July 2021). These, as well as other parameters (Earth Rotation Parameters (ERPs) and satellite phase biases), are necessary for subsequent users to process accurate station positions and tropospheric delays for their own applications (e.g., geodesy, surveying, weather forecasting, etc.). Both the IGS and user data processing procedures rely on the use of various models to account for geophysical effects (solid Earth tides, ocean tides), atmospheric propagation effects through the ionosphere and the neutral atmosphere, and instrumental biases and variations induced by the transmitter and receiver antennas and electronics [24]. The propagation effect in the neutral atmosphere, so-called tropospheric delay, is traditionally decomposed into its hydrostatic and wet components

in the zenith direction, which are mapped into the slant observation direction which their respective mapping functions, and a two-parameter horizontal gradient model [25]:

$$STD = ZHD \times m_h(\epsilon) + ZWD \times m_w(\epsilon) + (\vec{G} \cdot \vec{e}) \times m_G(\epsilon), \quad (1)$$

where ZHD is the Zenith Hydrostatic Delay (ZHD), mainly due to the contribution from dry air, ZWD is the Zenith Wet Delay (ZWD), due to water vapor molecules, $\vec{G}$ is the horizontal gradient vector, $\vec{e}$ is the unit vector pointing into the direction of the observed satellite, m_h, m_w, and m_G are the respective mapping functions, and ϵ is the elevation angle.

During the GNSS data processing, ZHD is corrected a priori, while ZWD and G are estimated. The ZWD estimates can include some residual bias from incorrect a priori ZHD correction or a deficiency of the hydrostatic and wet mapping functions. The GNSS software traditionally provide the total Zenith Tropospheric Delay (ZTD), which is the sum of the a priori ZHD, ZHD_{ap}, and estimated ZWD, ZWD_{est}:

$$ZTD = ZHD_{ap} + ZWD_{est}. \quad (2)$$

The main parameter of interest for climate studies is the Integrated Water Vapor (IWV), i.e., the integral of the water vapor molecules at the zenith. The conversion of ZTD into IWV operates, thus, in two steps [25]:

$$ZWD = ZTD - ZHD \quad \text{and} \quad IWV = \kappa(T_m) \times ZWD. \quad (3)$$

The accuracy of the IWV estimates is, thus, basically determined by the accuracy of the ZTD parameters derived from the GNSS data processing and the quality of the ZWD to IWV conversion procedure. For the purpose of retrieving IWV, a more accurate estimate of ZHD is required than the one used a priori for the processing. It can be derived from the surface air pressure data, P_s, available at the GNSS station [26]:

$$ZHD = 10^{-6} k_1 R_d \frac{P_s}{g_m}, \quad (4)$$

where k_1 is the dry air refractivity coefficient, R_d is the dry air specific gas constant, P_s is the surface air pressure, and g_m is the mean acceleration due to gravity. The ZWD to IWV conversion factor κ is defined as a function of the weighted mean temperature T_m [10]:

$$\kappa(T_m) = \frac{10^6}{R_v(k'_2 + \frac{k_3}{T_m})},$$

where k'_2 and k_3 are refractivity coefficients for the water molecule, and R_v is the specific gas constant for water vapor. The refractivity coefficients were taken from Bock [27]. The weighted mean temperature is defined by Bevis et al. [10]:

$$T_m = \frac{\int \rho_v(z) dz}{\int \frac{\rho_v(z)}{T(z)} dz}, \quad (5)$$

where $\rho_v(z)$ and $T(z)$ are the specific mass of water vapor and the air temperature, respectively, at height z above the surface. The integral is from the surface to the top of the atmosphere. It can be computed from a vertical profile of $\rho_v(z)$ and $T(z)$ given by a radiosonde climatology or an atmospheric model.

In this study, the auxiliary data, P_s and T_m, required for the conversion of the ZTD estimates into IWV are computed from a global atmospheric reanalysis. Using reanalysis data has several advantages: the data are available at any position and time on the globe, the pressure and temperature are well constrained by observations, making the reanalysis data the best estimate of the global atmospheric state at any position and time, and the assimilation system uses an efficient screening and bias correction procedure to reject

suspect observational data and adjust bias changes associated to observational system changes (e.g., between older and new satellites). In this study, we will consider two different reanalyses: ERA-Interim [15] and ERA5 [18]. Details on the reanalyses are given in the subsection below. The P_s and T_m data are computed from 6-hourly pressure level data, as described in Reference [14] and the subsequent IWV estimates are aggregated into daily estimates, as described in Reference [17]. The interest of comparing two different reanalysis is that ERA5 is of superior quality due to the assimilation of more observations and of higher spatial resolution, i.e., providing more accurate pressure and temperature estimates in regions of steep topography. The accuracy of reanalysis estimates of P_s and T_m, as compared to other data sources, is further discussed in Bock [27] and Bock et al. [28].

The main factors conditioning the accuracy of the ZTD estimates at the GNSS data processing level are, by decreasing order, the hydrostatic and wet mapping functions, the a priori ZHD correction data, and the antenna phase center variation (PCV) models. Any bias in these data and models would map directly into the ZWD estimates. The mapping functions are the most important because they determine how the signals from the satellites at various elevations are mapped into the zenith direction (from a mathematical point of view, they represent the partial derivatives, or regressors, of the ZHD and ZWD parameters). Because the hydrostatic delay is corrected a priori, any bias in the a priori ZHD, or in the hydrostatic mapping function, will also map into the estimated ZWD [29]. Receiver antenna phase center variations also highly depend on the elevation angle and, to a lesser extent, on the azimuth angle. Before 2005, relative calibration models were used in the IGS network, which were progressively replaced with absolute calibrations [30]. The convention at IGS is to use a type-mean calibration, i.e., a mean calibration model determined from several calibrated antennas of the same type (producer and product) when available, and a specific variant for each antenna/radome combination. When no specific antenna/radome combination exists, the rule is to "adopt" the antenna calibration without radome. In addition, when the antenna calibration does not exist, the calibration is "copied" from a similar antenna (based on electronic and mechanic properties). The official antenna/radome calibrations are distributed by IGS in the form of ANTEX (Antenna Exchange Format) files, e.g., igs05.atx and igs08.atx, at the times when the IGS and CODE data sets used in this study were produced. Absolute calibrations from "Robot" and "Chamber" are the most accurate, while relative calibrations (type "field") and relative converted to absolute (type "converted") are the less accurate [31]. The impact of satellite and receiver antenna offsets (PCO) and PCVs on geodetic parameters have been mainly investigated for positioning purposes. The impact on ZTD estimates has not been much studied yet, although it is known that it would be tightly correlated with the vertical position component. One of the goals of this paper is, thus, to examine how the change from igs05.atx to igs08.atx impacts the accuracy and homogeneity of the GNSS IWV series and to which extent these differences are detected by our segmentation method. In addition to the aforementioned factors, there are other errors sources that can impact the accuracy of the ZTD estimates, such as multipath and ambiguity fixing errors, as well as satellite orbits and clock errors, and unmodeled and mis-modeled station displacements at sub-daily time scales (e.g., tides). However, these are minor errors sources for the purpose of our study here. For further details, see the study of Ning et al. [32]. Of main concern here are the sources of bias and the mechanisms through which these biases can change with time, i.e., translate into inhomogeneities.

In this study, we consider two different GNSS data sets, which are representative of two different generations of reprocessing products delivered by IGS (see Table 1). The first one, referred to as IGS repro1, was produced by JPL/NASA in 2010–2011 in precise point positioning (PPP) mode with GIPSY OASIS II software [33] as a special release of ZTD estimates. This data set used the reprocessed IGS orbits, clocks, and ERPs produced by JPL/NASA in the framework of the 1st reprocessing campaign organized by IGS. The reprocessed satellite products were generated for the period 1 January 1995 to 31 December 2007, but JPL completed the series until mid-2011 in a consistent way. In this study, we use the

ZTD estimates until 31 December 2010 to have an integer number of years. According to the discussion above, the prominent features of the processing procedure are:

- Standard Temperature and Pressure (STP) model used for a priori ZHD correction [34],
- Global mapping function (GMF) for the hydrostatic and wet delays [35],
- IGS05 reference frame and igs05.atx absolute antenna PCO/PCV models
- 7° elevation cutoff angle, and
- uniform observation weighting.

Table 1. Processing strategies used by JPL/NASA to produce the IGS first reprocessed tropospheric data set (IGS repro1) and by the Center for Orbit Determination in Europe (CODE) to produce the CODE REPRO2015 reprocessed data set.

	IGS Repro1	**CODE REPRO2015**
Software	GIPSY OASIS II	Bernese GNSS software v5.3
Strategy	PPP solution	Double-difference solution of a global network
Orbits, clocks, ERPs	IGS repro1 (1995.0–2008.0) + IGS final (2008.0–2011.0)	CODE repro2 (1994.0–2015.0) + CODE final (2015.0–2019.0)
Reference frame	IGS05	IGb08
Antenna calibration	igs05.atx	igs08_1852.atx until 28 January 2017, igs14.atx from 29 January 2017
Window length	24 h	72 h
Elevation cutoff angle	7°	3°
Observations	GPS	GPS (1994.0–2002.0), GPS+GLONASS (2002.0–2019.0)
Observation sampling	5 min	3 min
Observation weighting	uniform	$\sigma^2 = 36 \times 10^{-6}/\cos^2(Z)$ where Z = zenith angle
Tropospheric model	ZHD and ZWD a priori: ZHD = 1.013 × 2.27 × exp(−0.116*ht), ZWD = 0.1 m. GMF mapping functions (hydrostatic and wet). Random Walk model for ZWD and gradient parameters with constraints: 3 mm $h^{-1/2}$ (ZWD) and 0.3 mm $h^{-1/2}$ (gradients). ZWD and gradient sampling: 5 min	ZHD and ZWD a priori: 6-hourly ECMWF analysis (provided by TUV). VMF1 mapping functions (hydrostatic and wet). Piece-wise linear model for ZWD with constraints: 5 m absolute and 5 m relative. Sampling : 2 h (ZWD), 24 h (gradients).
Tropo files	ZTD and gradient estimates provided in SINEX files (0000, 0005, . . . 2345 UTC)	ZTD and gradient estimates provided in SINEX files (resampled to 01, 03, . . . 23 UTC)
Coordinate estimates	Estimated once per 24 h	Estimated once per 24 h
Ambiguity resolution	Float	Fixed

It should be noted that, due to a flaw in the data file handling, the ZTD files for a number of stations are not available on the FTP repository, and older files were used instead, for which a major difference was the use of the older NMF hydrostatic and wet mapping functions [36].

The second GNSS data set, referred to as CODE REPRO2015, was processed by the Center for Orbit Determination in Europe (CODE) in 2015 [37] using the Bernese GNSS software [38] in network mode (double-differenced observations). This data set used the reprocessed IGS orbits, clocks, and ERPs produced by CODE in the framework of the 2nd reprocessing campaign organized by IGS. The reprocessed products cover the period from 1 January 1994 to 31 December 2014. For the purpose of this study, we completed the series with the operational CODE products until the end of 2018 [39]. However, the latter underwent a switch in the reference frame to IGS14 and in the antenna cnetwork alibration (double-differenced observations).model to igs14.atx, on 29 January 2017. This change was, thus, included in the metadata for the CODE solution.

The prominent features of the CODE processing procedure are:

- ECWMF grid estimates of a priori ZHD and ZWD distributed by Technical University of Vienna [40],
- Vienna mapping function (VMF1) for the hydrostatic and wet delays [40],
- IGb08 reference frame and igs08_1852.atx absolute antenna PCO/PCV models,
- 3° elevation cutoff angle,
- $1/\cos(\text{zenith})^2$ observation weighting, and
- GPS observations (1994–2001) and GPS + GLONASS (2002–2014).

According to the discussion above, we see that all the main factors conditioning the accuracy of the ZTD estimates are different between the two data sets. Not discussed above are the elevation cutoff angle and observation weighting, which contribute significantly to the sensitivity of ZTD estimates to the various elevation-dependent error sources (mapping functions, antenna PCO/PCV models, and multipath). The purpose of using a lower elevation cutoff angle in CODE is to include more observations, i.e., increase the precision of the estimated parameters. However, multipath is generally higher at low elevations. To mitigate it, the lower elevation observations are down-weighted. The JPL/NASA processing strategy was different as they used a 7° cutoff angle and no down-weighting. Possibly, this strategy might be more sensitive to multipath and anomalous propagation effects at low elevations.

The ZTD estimates from both data sets were screened for outliers following the methodology described in Bock [41] and Stepniak et al. [42], and converted to IWV using either ERA-Interim or ERA5 reanalysis. The 6-hourly IWV data were compared to the reanalysis IWV data and further screened for outliers (for each station, the IWV differences exceeding the median ± five standard deviations were removed). Afterward, the IWV values from GNSS and reanalysis, and the IWV differences between GNSS and reanalysis, were aggregated into daily and monthly estimates and made publicly available on the AERIS data center [43,44].

Figure 1 shows the location of the GNSS station available from the two data sets. In this study, we selected 81 common stations, for which the time series in both data sets covered a period of at least 15 years.

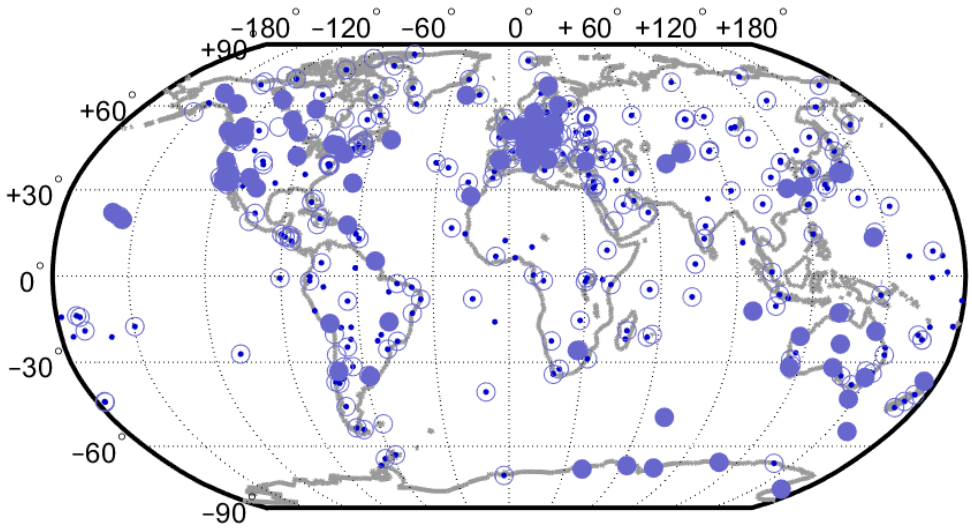


Figure 1. Map of the GNSS stations available from the two reprocessed data sets: IGS repro1 (empty circles), CODE REPRO2015 (small dots), and the 81 common stations (full circles) used in this study.

2.2. Reference IWV Data

Our homogenization method operates on IWV differences between a GNSS series and a reference series. Because the IGS network is quite sparse, we cannot use a nearby station as is commonly done by climatologists (as in Venema et al. [9]). Instead, for every

GNSS station, a series of IWV from each of the two reanalyses is derived, and daily IWV differences are formed, as explained above. In earlier studies, we found that ERA-Interim and GNSS IWV had significant representativeness differences in Antarctica and in regions of steep topography (Andes, Himalayas, etc.) or near the oceans [17]. In this study, we will investigate the impact of representativeness errors on the segmentation results by comparing the results from the two reanalyses. The spatial resolution of the reanalyses is 0.75° × 0.75° for ERA-Interim and 0.25° × 0.25° for ERA5. Reduced representativeness errors are, thus, expected from ERA5 data. Moreover, the IWV values computed from ERA5 are also expected to be of higher quality since this reanalysis used a more recent model and assimilation system, and assimilated a much larger number of observations, especially in recent years [18].

2.3. Homogenization Method

Figure 2 shows the data flow chart starting with the GNSS ZTD data and ending with the corrected IWV series. The first two steps (Conversion and Comparison) are described in the previous sub-section.

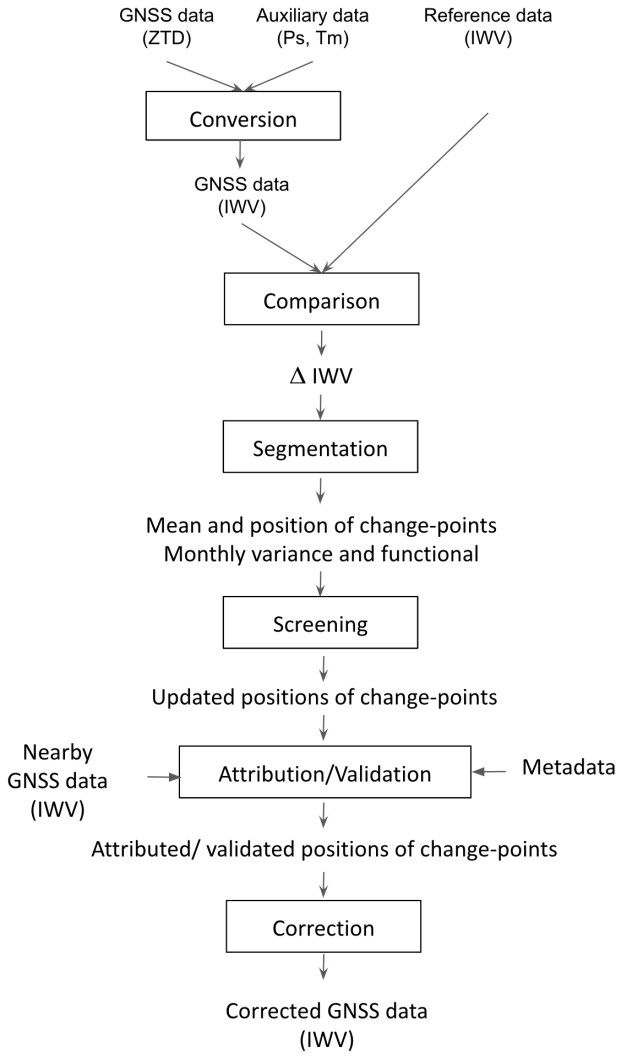


Figure 2. Flowchart of the general homogenization procedure.

The third step is the segmentation, i.e., the detection of change-points in the mean of the IWV difference time series. Here, we use the fast version of the GNSSfast R package published by Quarello et al. [45]. This version is available on https://github.com/arq16/GNSSfast.git, accessed on 30 July 2021. The segmentation method estimates K, the

number of segments, or the number of change-points K-1, the K means of the segments, and the K − 1 positions of the change-points, as well as a periodic bias function f and the monthly variance of the noise (also with annual periodicity), in a unified model for each IWV difference time series. It has been shown in a previous paper that modeling a periodic bias function and a heterogeneous noise helps to accommodate for the representativeness differences between the GNSS IWV data and the reanalysis IWV data (see Quarello [21]). The inference procedure is based on a penalized likelihood approach. Several penalty criteria are considered, but here we will present only the results for the so-called Birgé and Massart's penalty based on the 'dimension jump' proposed by Birgé and Massart [46] and calibrated in Lebarbier [47]. In a previous study, the GNSSseg method was compared to other existing segmentation methods on the basis of a benchmark data set and was found to be one of the most efficient [22].

The noise in the IWV differences is generally not perfectly Gaussian and strong peaks can lead to spurious change-points. These false detections are checked and screened out by a special post-processing procedure, which is symbolized in Figure 2 by the Screening step. In this step, clusters of 2 or more nearby change-points are checked for a significant change in the mean before and after using a weighted t-test. In this study, we consider as a cluster all the change-points closer than 80 days (the value of 80 days was found as optimal based on a mixture model analysis). The change-points within a cluster are flagged as "outliers" in the following. If the change in the mean is not significant (at the level of 5%), all the change-points in the cluster are removed. If it is significant, the group of change-points is replaced with one change-point, and its position is taken as the mean of the positions.

With a relative segmentation approach, such as used here, climatologists often assume that the reference is homogeneous [9]. If the reference is not homogeneous, multiple comparisons are necessary to determine in a statistical way to which of the series each detected change-points belong to. This task is accomplished during the Attribution/Validation step. The attribution of a change-point to GNSS or reference is decided based on the comparison with nearby GNSS series, while the validation is referring to the comparison of the detected change-points with the GNSS metadata. For the attribution step, the idea is to use all the available stations in the GNSS data, as long as the time series covers a period of time encompassing the tested change-point. We tested the procedure with the CODE REPRO2015 data set, including all 434 stations, with a maximum distance of 200 km and a time window of ±six months around the change-points. Based on the available data, about 30% of the change-points from the selected 81 stations could be checked. This number is too small to apply this test in a systematic way to all the data sets used in this study. So, we decided not to include it in the general discussion but only for a few cases studies. Instead, we will only use the validation step with respect to the GNSS metadata with a window of ±62 days, as in Van Malderen et al. [22].

The last step is the correction of the GNSS time series for the changes in the mean based on the attributed/validated change-points. Here, we follow the approach of Van Malderen et al. [22] which consists in subtracting from the GNSS IWV series a piece-wise constant function constructed from the change in means of the GNSS-reference series, where the last segment is taken as a reference. The performance of the correction depends on all the previous steps.

2.4. Trend Estimation Method

Following Weatherhead et al. [48], we use a linear trend model:

$$Y_t = \mu + \omega X_t + S_t + N_t, \tag{6}$$

where Y_t is the IWV time series, X_t the linear trend function, S_t a seasonal component which will be represented by a fourth order Fourier Series, $S_t = \sum_{i=1}^{4} a_i \cos(2\pi it/T) + b_i \sin(2\pi it/T)$, t is time in days since some reference date, $T = 365.25$ days, and N_t the noise which is assumed to be autoregressive of the order 1 (AR(1)), that is, $N_t = \phi N_{t-1} + \epsilon_t$, where the ϵ_t are independent random variables with zero mean and variance σ_ϵ^2. The AR(1)

noise is a good statistical representation of the day-to-day variability in the IWV time series around the mean seasonal cycle. The unknown parameters of this model are: μ, the mean IWV, ω the slope of the linear trend, the a_i and b_i coefficients of the Fourier Series, ϕ the autocorrelation of the noise, and σ_ϵ^2 the variance of the noise. We use a Generalized Least Squares (GLS) algorithm to estimate all the parameters and their formal errors.

3. Results

3.1. Segmentation Results

This section discusses the segmentation results and how they are impacted by four factors: (1) the GNSS data processing (IGS versus CODE), (2) the length of time series (short period, 1994–2010, and long period, 1994–2018), (3) the reference data set (ERAI versus ERA5), and (4) the auxiliary data used in the ZTD to IWV conversion (ERAI versus ERA5).

Table 2 summarizes the segmentation results for all four comparisons. Statistics are given for 81 common stations in both GNSS data sets. They include average data properties, such as the mean of the estimated monthly variances and the standard deviation of the estimated functional (stdf). The segmentation results are compared by means of the total number of detected change-points, both before and after the screening, the number of outliers, the number of metadata validations, and the number of similar detections.

Table 2. Summary of pairwise comparisons of segmentation results from various data sets used in this work. The validation with respect to GNSS metadata and the similar detection statistics used a closeness window of ±62 days. (a) Segmentation is run over the full time series (1994–2018), but change-points are compared for the time-limited period (1994–2010). (b) This CODE data set uses auxiliary data from ERA5. (c) This CODE data set uses auxiliary data from ERAI.

	(1) Impact of Processing		**(2) Impact of Time Length**		**(3) Impact of Reference**		**(4) Impact of Auxiliary**	
Data Set	IGS—ERAI Time-Matched	CODE—ERAI Time-Matched	CODE—ERAI Time-Limited	CODE —ERAI	CODE (b) —ERAI	CODE (b) —ERA5	CODE (b) —ERA5	CODE (c) —ERA5
Time span	1995–2010	1995–2010	1994–2010	1994–2018 (a)	1994–2018	1994–2018	1994–2018	1994–2018
Mean of the monthly variances (kg m^{-2})	0.68	0.62	0.62	0.63	0.61	0.46	0.46	0.46
Standard deviation of the functional (kg m^{-2})	0.26	0.24	0.24	0.23	0.23	0.17	0.17	0.17
No. detections	231	257	296	249	364	398	398	392
No. outliers	36	38	73	40	60	71	71	87
No. detections after screening	211	235	252	227	333	359	359	343
Validations after screening	63	68	77	78	114	131	131	125
Validations after screening (%)	29.9	28.9	30.6	34.4	34.2	36.5	36.5	36.4
Similar detections	103∼48.8%		185∼81.5%		151 ∼45.3%		243 ∼70.9%	

3.1.1. Impact of GNSS Data Processing

In comparison to CODE, IGS is noisier with respect to the reanalysis with about 10% higher monthly variances in the time-matched data on average (Table 2). Inspection of the monthly variances by station shows that the noise is systematically larger in the IGS data set for all stations except at USUD (Figure 3c). All the stations located in Antarctica (DAV1, MAW1, SYOG, MCM4, MAC1, CAS1) show a significantly larger relative monthly variance in the IGS data set, with a maximum of 43% at DAV1. We believe that the use of the VMF1 mapping function in the CODE solution makes a major difference as it accounts for the high-resolution (6-hourly) temporal variations in the atmospheric (dry and wet) layering, whereas the GMF used in the IGS solution only models a smooth seasonal variation. On the other hand, the stdf, which quantifies the magnitude of the periodic bias, does not show such a systematic difference, although CODE shows smaller values again on average. A majority of stations in the IGS data set have slightly larger stdf (54 stations, i.e., 67%,

larger versus 27 stations, i.e., 33%, smaller). For a few stations, the difference in stdf is relatively big, exceeding ±50% (Figure 3d,e). The difference in stdf can be explained by the difference in mapping functions, a priori ZHD, and elevation cutoff angle. Specific cases are discussed below.

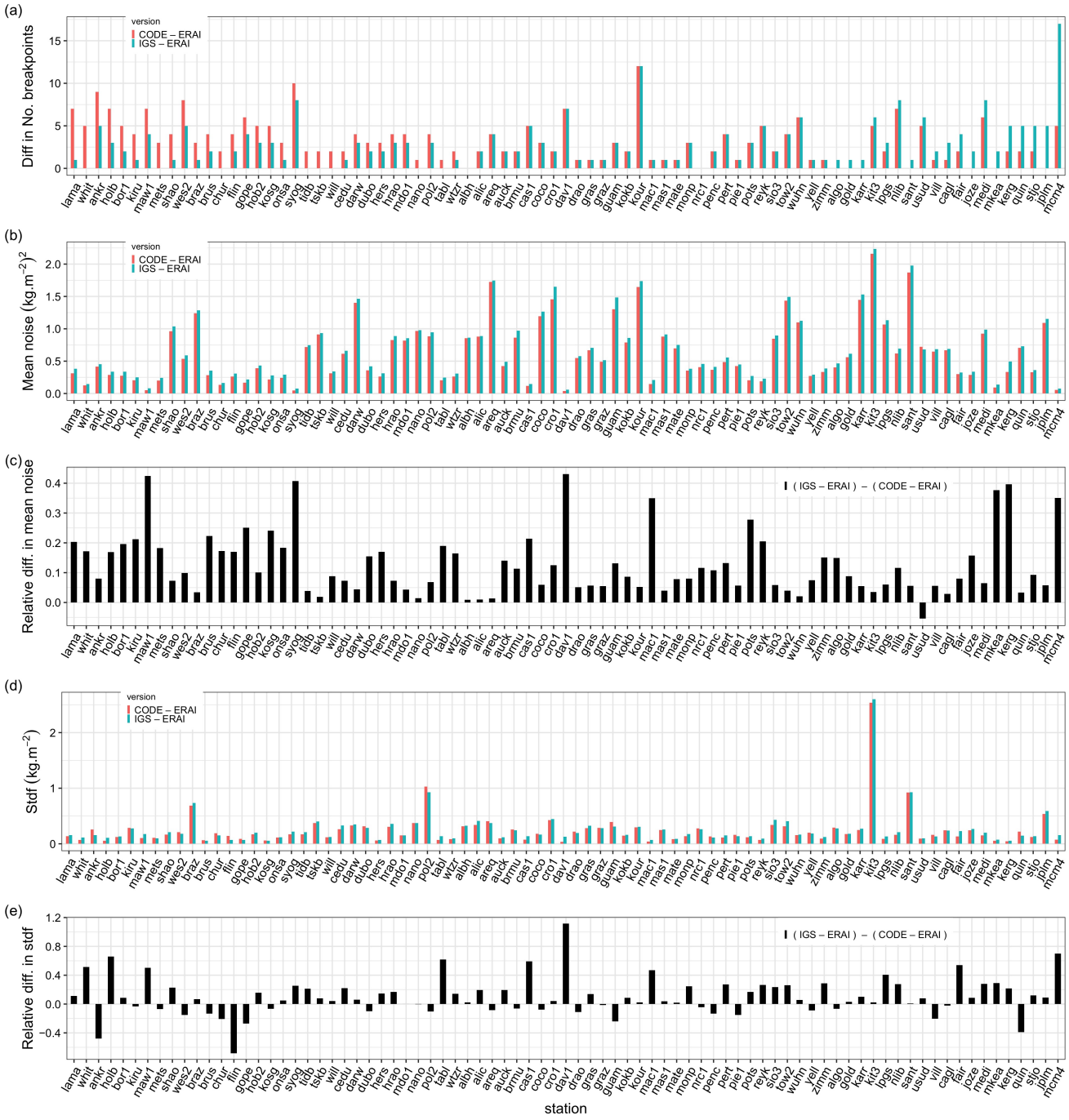


Figure 3. Comparison of segmentation results from two different GNSS data sets, IGS repro1 (IGS), and CODE REPRO2015 (CODE), at 81 common stations: (**a**) number of detected change-points, (**b**) mean of the monthly variance, (**c**) relative difference in the mean of monthly variance (IGS − CODE), (**d**) standard deviation of the functional modeling the periodic bias (stdf), (**e**) relative difference in the standard deviation of the functional (IGS − CODE). All the parameters are estimated from time-matched series for the period 1995–2010. The relative differences are computed relative to the mean (IGS + CODE)/2.

The segmentation found more change-points in the CODE data set, both before and after the screening, compared to IGS, and slightly more outliers (38 versus 36). After screening, the validation percentages of the two data sets are very similar (28.9 and 29.9%).

However, the number of similar change-points is relatively small (48.8%), indicating a strong dependence of the segmentation results on the GNSS processing method. There are several reasons that can explain the difference in the number, as well as the position, of change-points. Firstly, the segmentation is sensitive to the magnitude and stationarity of the noise. Since the noise is higher for the IGS data set, a number of small offsets are not detected there. Secondly, IGS includes mapping function changes in 2008 and 2009, which are not included in CODE. They lead to 15 extra-detections in IGS. Thirdly, the improved tropospheric model and updated antenna/radome calibrations used in the CODE processing are likely to reduce some mean biases and seasonal variations in addition to the reduced noise, which may lead to a difference in the number and positions of change-points (e.g., higher periodic biases in IGS may lead to extra change-points). To summarize, some of the data features may increase the number of change-points in IGS, while others may increase the number in CODE. Below, we examine a few special cases.

Figure 4 shows a pathological example (station MCMC4, McMurdo, Antarctica). For this station and all other stations in Antarctica, the difference in a priori ZHD leads to a reduced bias (mean and seasonal) in the CODE IWV estimates with respect to ERAI. The mean difference between CODE and IGS for MCM4 is about -0.5 kg m^{-2}. Large oscillations are seen in both solutions during the period 2002–2006. These oscillations are believed to arise in the GNSS measurements due to snow intrusion between the antenna and radome as discussed by Koulali and Clarke [49]. Most of the stations in Antarctica show this feature for some periods. In the CODE solution, the oscillations are slightly smaller, which may be explained by the use of a lower cutoff angle which would mitigate the anomalous path delay variations partly. In the IGS series, these oscillations lead to several extra change-points, although they are not due to equipment changes and, thus, are not validated by the metadata. The mapping function change in 2008 (from GMF to NMF) and 2009 (from NMF to GMF) in IGS also leads to 2 additional change-points. As a consequence, the IGS series at MCM4 has 17 change-points, including 6 outliers due to high noise, whereas the CODE solution has only 5 change-points. The number of validations with respect to metadata are 5 and 2, for IGS and CODE, respectively, with one validated change-point in 2006 (due to a receiver change) similar in both cases.

Next, we examine another example (station GOPE, Czech Republic), where the equipment changes have a strong and different impact on the CODE and IGS solutions. Over the study period (1995–2010), many changes occurred at this station. Figure 5 gives a visual description of these changes based on the IGS metadata, where each color represents a specific equipment type (producer and product), and the vertical dashed lines indicate a sub-type change (serial number or firmware for receivers). Five different antenna/radome types were used at this site: TRM14532.00/NONE (from 13 May 1995 to 4 November 1999 and from 24 July 2000 to 4 October 2000), ASH701073.3/SNOW (from 4 November 1999 to 24 July 2000), ASH701946.3/NONE (from 4 October 2000 to 14 July 2006), and TPSCR3_GGD/CONE (from 14 July 2006 to 14 December 2009), TPSCR.G3/TPSH (since 14 December 2009). Three different receiver types were used: TRIMBLE_4000SSE (from 13 May 1995 to 4 November 1999, and 24 July 2000 to 4 October 2000), ASHTECH_Z18 (from 4 November 1999 to 24 July 2000, and 4 October 2000 to 14 December 2009), and TPS_NETG3 (from 14 December 2009 to 8 August 2018). The figure shows that the same receivers were removed and reinstalled for some periods, as well as that they got several firmware updates (see the dashed vertical lines). The cutoff angle was also changed twice, going from 15 degrees to 5 degrees on 4 November 1999, and from 5 degrees down to 0 degree on 14 December 2009. The impact of these equipment and setting changes on the GPS-ERAI IWV differences is obvious from Figure 6. For example, there is a strong change in the mean IWV difference for both solutions before and after 4 October 2000. This change is detected (and validated) by the segmentation in both data sets. Before that date, CODE has 2 other change-points, while IGS has only one, although IGS shows similar changes in the mean to CODE but they are not detected. Neither of these change-points is explained by the metadata. After 4 October 2000, CODE has 3 more change-points, while IGS has only 2.

All 3 are validated in the case of CODE, and none is validated in the case of IGS. However, the IGS detected change-points seem correct as they well represent the changes in the mean of that series. The difference in the number and positions of change-points between the IGS and CODE segmentation results is actually due to the difference in the two GNSS IWV data, as highlighted in the lower plot of Figure 6. This difference can be explained by the different antenna/radome calibration models used in the two GNSS processing. The most striking bias between the two series is for the period from 14 July 2006 to 14 December 2009 where the TPSCR3_GGD/CONE antenna/radome pair is used. In the igs05.atx calibration file used in the IGS solution, this calibration was of "Field" type (i.e., an older relative calibration dating back to April 2005), while the CODE solution used a more recent "Robot" type calibration (from April 2013). For all other antenna/radome types used at this station, both GNSS solutions used similar calibration types ("Robot" or "Copied"), although the CODE calibrations were more recent in all cases, which may explain the small residual biases between the two solutions. Another abrupt shift in the CODE−IGS series is observed on 1 January 2002 which coincides with the introduction of GLONASS observations in the CODE solution. A closer look reveals that, for some reason, the shift is merely in the IGS−ERAI series and that the CODE−ERAI series is, rather, drifting.

The GOPE example shows that antenna/radome changes at a site are likely to produce an abrupt shift in the mean IWV, even though calibration models are used during the data processing. Different calibration types and versions have different biases, leading to different inhomogeneities in the two reprocessed GNSS data sets analyzed in this work and to different segmentation results. Many such cases could be found by inspecting the time series and segmentation results. In some cases, receiver changes and observation cutoff angle changes also induce abrupt shifts of different amplitudes for the two data sets. This feature might be due to the use of different processing cutoff angles (3 degrees in CODE and 7 degrees in IGS). In other cases, we can see a small drift (below 0.5 kg m^{-2}) between the two solutions, which might be due to the change in the number of satellites used in the processing, especially with the inclusion of GLONASS observations in the CODE solution starting from 2002, although not many GNSS stations were recording GLONASS data in the period from 2002 to 2010. Since the amplitude of such small drifts is generally small compared to the noise, it would not induce extra change-points in the segmentation; although this cannot be guaranteed for sure, it may impact the trend estimates.

3.1.2. Impact of the Length of Time Series

The impact of the length of time series is inspected from the comparison of CODE-ERAI segmentation results where the segmentation was run both on the full (1994–2018) series and on the series limited to 1994–2010. The change-points statistics (number of detections, outliers, and validations) reported in the second section of Table 2 are given for the common period (1994–2010).

It can be deduced from Table 2 that the estimated monthly mean variance and the estimated stdf are nearly similar on average over all stations. However, inspecting individual stations shows that small differences exist (Figure 7b–e). Slightly more than half of the stations have smaller variance in the full series (57 versus 43%) and smaller stdf (52 versus 48%). This difference can be explained by the fact that the GNSS IWV estimates become more accurate with time; thus, the agreement with the reanalysis improves in the more recent years included in the full series. On the other hand, the cases where the noise is larger in the full time series may either be due to situations where the GNSS measurements get noisier, or the ERA-Interim reanalysis becomes less accurate with time. This is, for example, the case at station COCO where the monthly variance is 25% higher in the full series. The same situation is observed at many tropical stations, although the quality of the GNSS data has generally improved at these sites. We suspect that the problem is in the quality of the ERAI reanalysis, which may indeed have degraded in recent years due to the end of several satellite missions [18]. One obvious case is with station COCO discussed in the next sub-section.

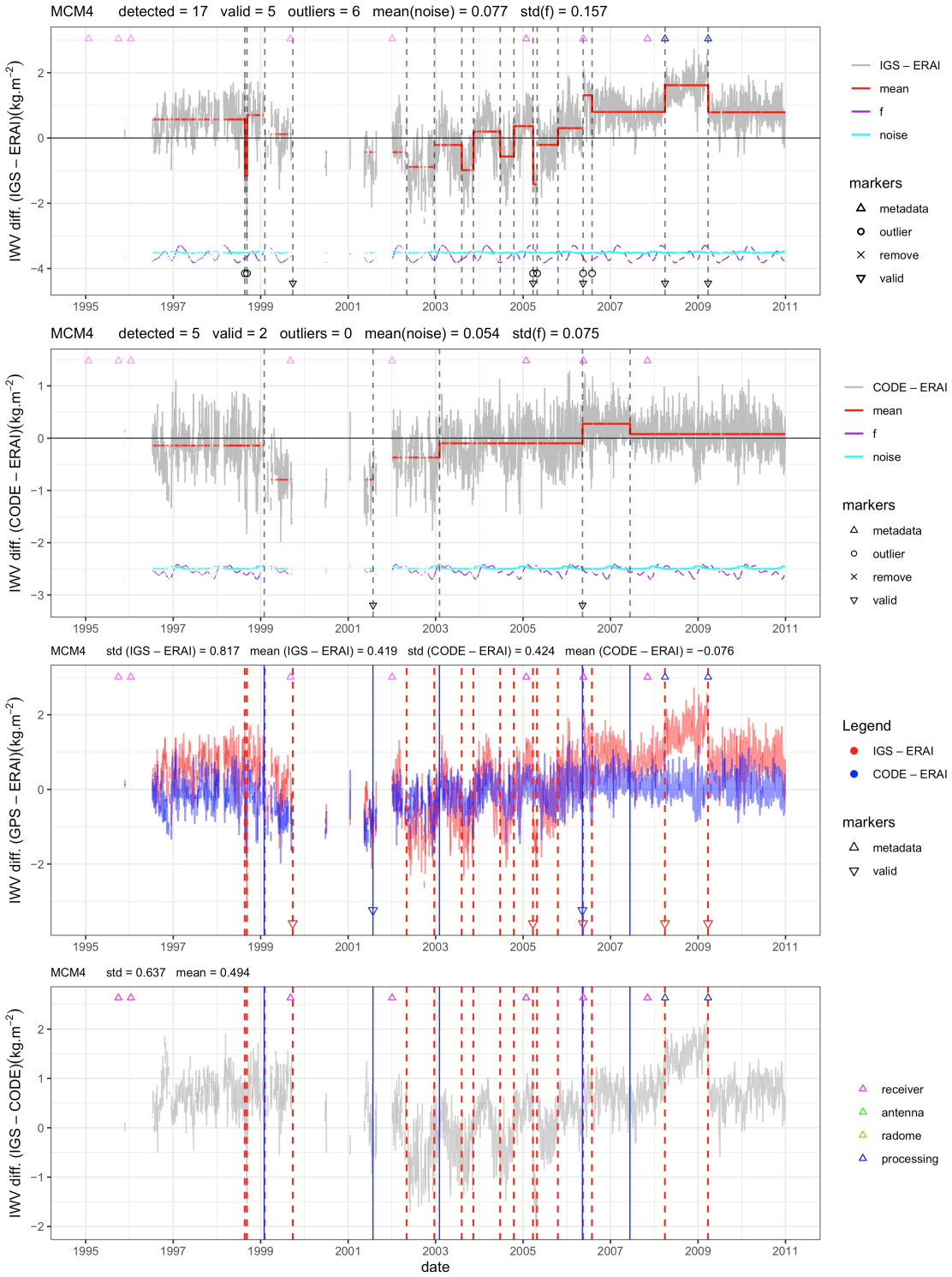


Figure 4. Time series of IWV differences at station MCM4 (McMurdo, Ross Island, Antarctica). The two top plots show the results for IGS-ERAI and CODE-ERAI, with segmentation results superposed (the red line shows the weighted mean as estimated by the segmentation method, where the black dashed lines indicate the change-point positions, the cyan line is the estimated monthly mean variance, and the purple line is the estimated functional modeling the periodic bias). The validated change-points are indicated by a marker at the lower end of the black dashed line. The 3rd plot from the top shows the two IWV difference series and detected change-points for clarity of the comparison. The lower plot shows the difference, IGS-CODE, with change-points superposed. In all plots, the markers at the top indicate the equipment changes. The color code is given in the lower plot. The processing changes apply only to IGS.

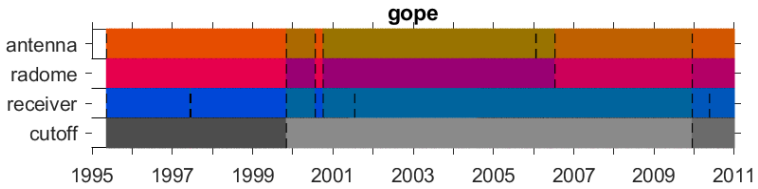


Figure 5. Schematic view of the equipment changes that occurred at station GOPE (Ondrejov, Czech Republic) during the period 1995–2010. Each color symbolizes a different equipment type (producer and product). Vertical dashed lines show additional changes with the same type (serial number or firmware upgrade in the case of receivers). Information extracted from the IGS site logs (ftp://igs.ign.fr/pub/igs/igscb/station/log/, accessed on 30 July 2021).

The number of change-points detected from the long time series in the common period is smaller than for the short time series, both before and after screening (Table 2). The same is found for the number of outliers, while the percentage of validations is consistently increased. These results indicate that, on average, the segmentation yields more accurate results with the long time series. The difference in the number of change-points can be understood from the fact that the penalty criterion is conservative and tends to choose a small number of change-points. For instance, the total number of change-points found from the long time series is 321 (not shown), i.e., not much larger than the total number for the short time series (296, as reported in Table 2). Because of this conservative property, the number and positions of the change-points in the two series cannot be expected to be the same at most stations. Figure 7 shows that 39 stations (48%) have the same number of change-points (although the positions may not be the same), 16 stations (20%) have a larger number in the long series, and 26 stations (32%) have a smaller number in the long series. Nevertheless, 80.8% of the change-points are similar (i.e., within $\pm$ 62 days).

Figure 8 shows the example of station VILL (Villafranca, Spain), where the mean noise and stdf are very similar for the two-time series, but the segmentation results are quite different: 6 change-points are detected in the short time series and 2 in the long time series. Two change-points are similar in both series and are validated. The additional change-points found in the short time series capture short but significant variations in the mean. These change-points are not retained in the long time series because other such variations are seen all along in the long time series. The penalty criterion avoids selecting all those change-points, and the final optimal solution eventually has only one change-point. This solution seems more reasonable.

3.1.3. Impact of the Reference Data Set

The third section of Table 2 summarizes the segmentation results when either ERAI or ERA5 is used as a reference, i.e., the segmentation is run for CODE-ERAI or CODE-ERA5 IWV differences, when the same auxiliary data is used in the GNSS ZTD to IWV conversion (here from ERA5). Globally, both the mean noise and the stdf with ERA5 are reduced by 25% compared to ERAI. Figure 9c,e show that the reduction in noise and stdf is observed at 95% and 75% of all stations, respectively. This difference can be explained by the lower representativeness error in ERA5 due to higher spatial resolution, as well as higher quality of the IWV temporal variations in this reanalysis, probably due to the assimilation of more satellite observations. Figure 10 shows the most striking case of increase in noise with time for ERAI. The impact on the segmentation results is quite significant. Only one change-point is detected in the more noisy series, while seven change-points are detected in the less noisy one. At many stations, there is also an excess noise in ERAI during the moist period of each year; see the example of station KIRU (Kiruna, Sweden) in Figure 11. At this station, the CODE-ERAI series has much larger seasonal variations in the noise and in the functional than CODE-ERA5. This leads to more change-points in more noisy series (6 versus 2) because of the sharp increase in the noise during some years, which is not well represented by the periodic bias and monthly variance. As a result, four outliers are detected in the CODE-ERAI segmentation results. In the CODE-ERA5, the two change-

points are validated by the metadata, which is, again, better than in the CODE-ERAI results. Other examples with large changes in stdf are stations KIT3, POL2, and SANT (Figure 9d). These stations were also pointed as extreme cases of representativeness errors in ERAI by Bock and Parracho [17] due to steep orography.

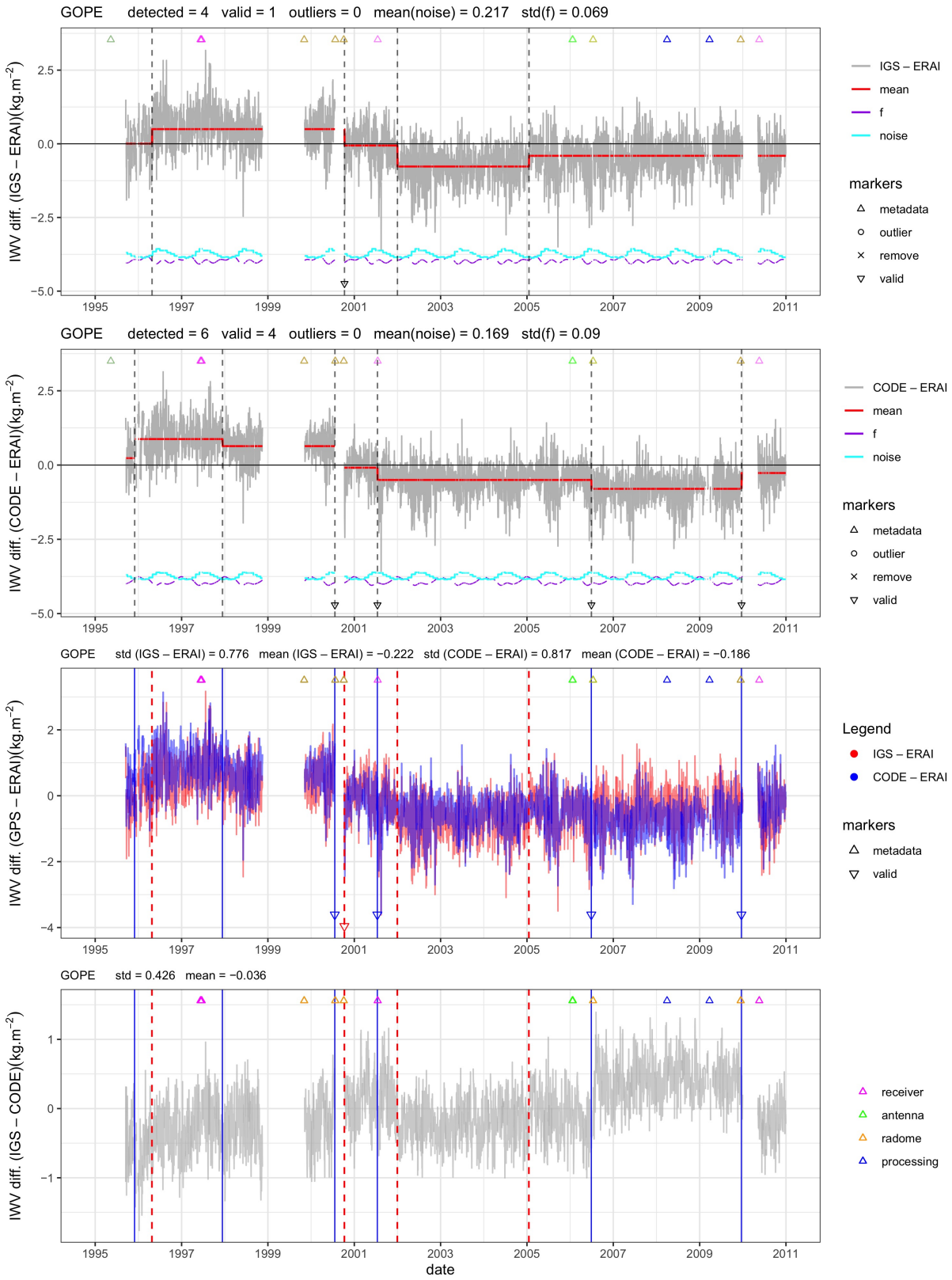


Figure 6. Similar to Figure 4, but for station GOPE (Ondrejov, Czech Republic).

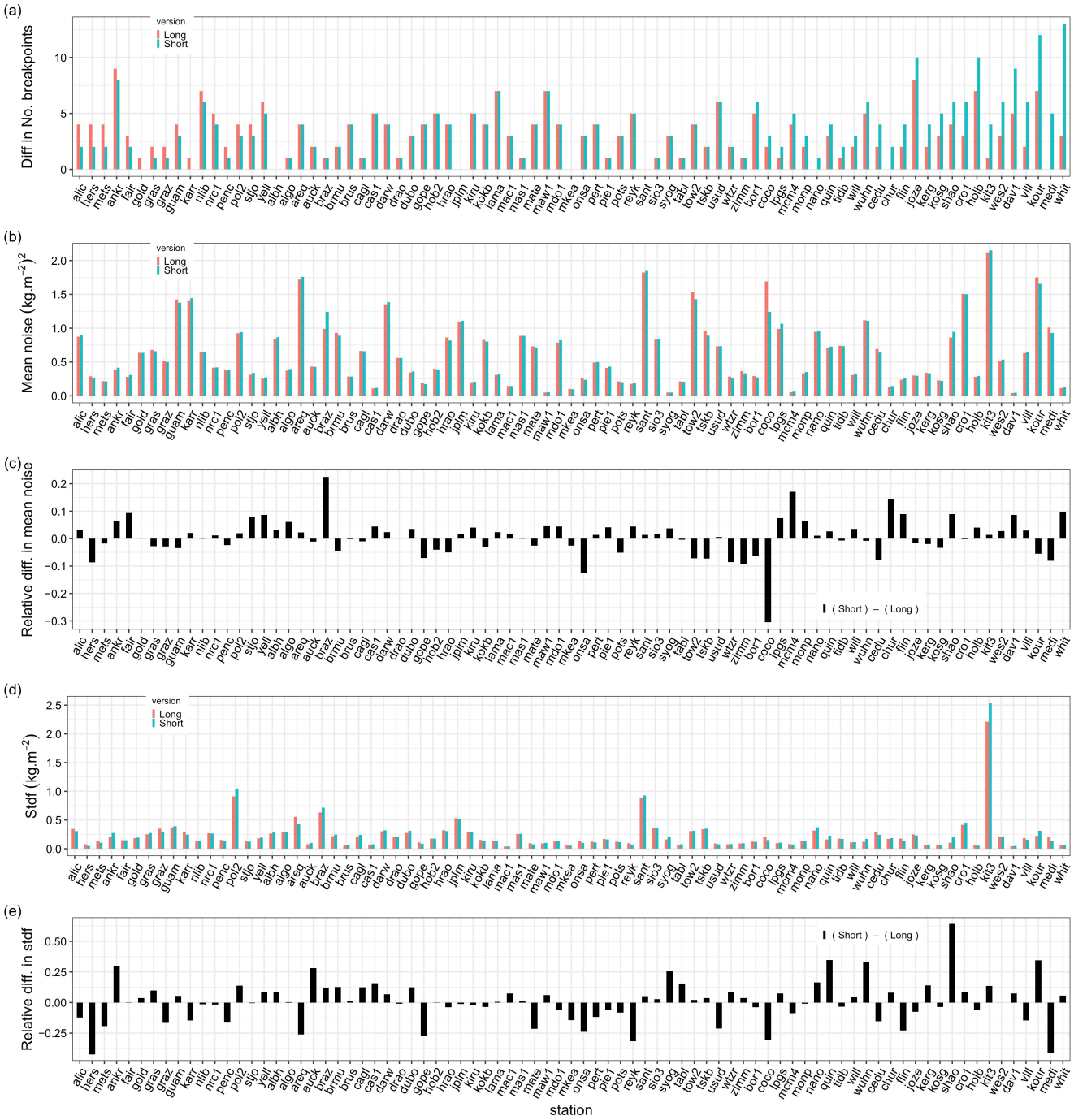


Figure 7. Similar to Figure 3, but comparing segmentation results from two different lengths of the CODE data set (long time series, from 1994 to 2018, short time series, from 1994 to 2010). The mean noise and standard deviation of the functional (stdf) are representative of the full times series, but the number of breakpoints are counted in the common period only (1994–2010).

From Table 2, we see that the total number of change-points is larger when ERA5 is used as a reference. This can be understood as the consequence of the general decrease of the noise and periodic bias with this reanalysis, as discussed in the case of station COCO above. When the noise is decreased, it is easier to detect a small offset in the time series. With the increase in the number of change-points, the number of outliers is increased, as well. However, after screening, the percentage of validations is higher with ERA5 as a reference. So, there is a clear benefit of using the more recent reanalysis as a reference.

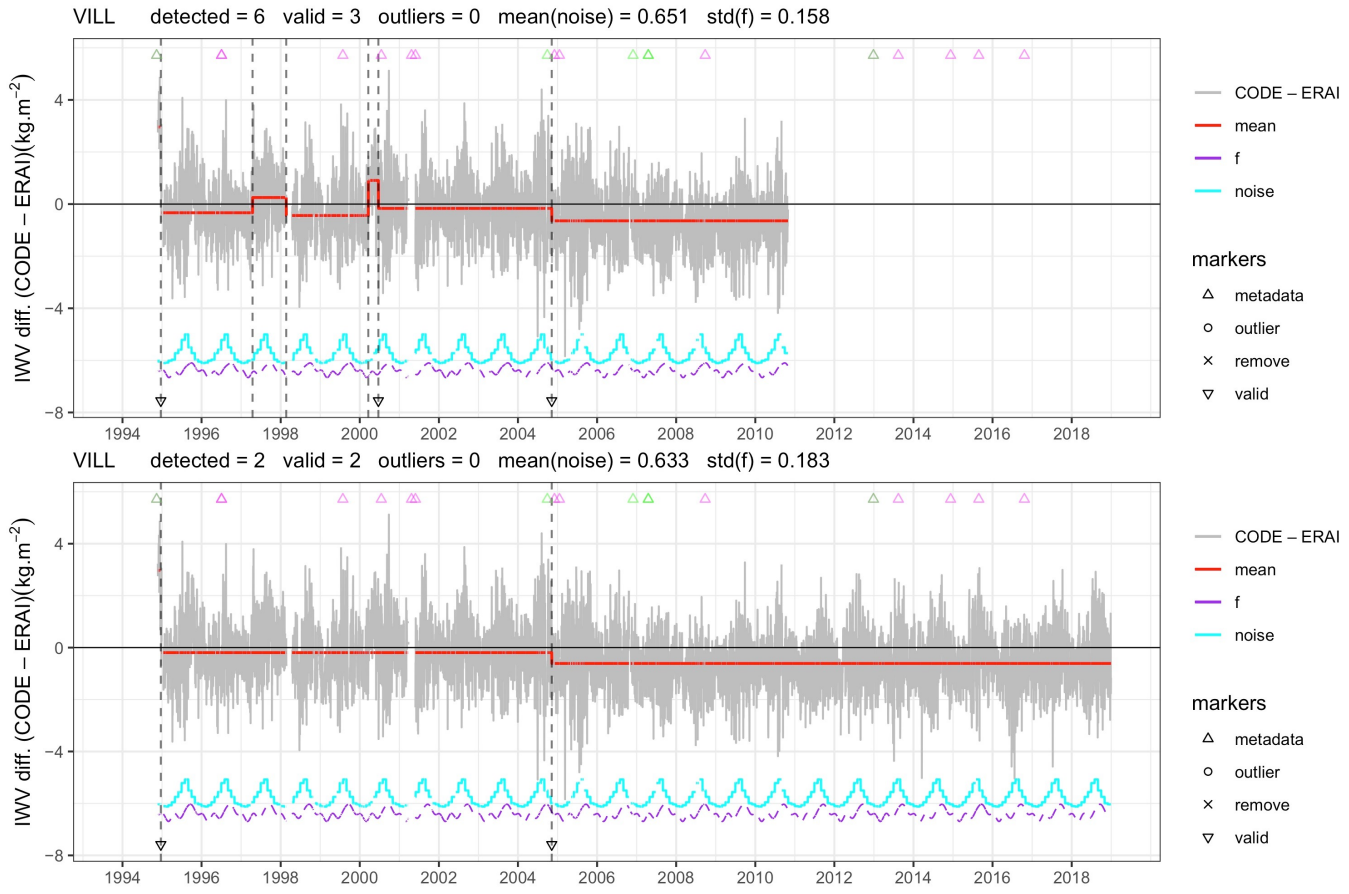


Figure 8. Similar to Figure 4, but for station VILL (Villafranca, Spain).

Figure 9a shows that 37 stations (46%) in CODE-ERA5 have a higher number of change-points than CODE-ERAI, 29 stations (36%) have a smaller number, and 15 stations have a similar number. From Figure 9c,e, we see that MKEA (Mauna Kea volcano,Mauna Kea , USA) and USUD (Usuda, Japan) are two cases where the mean noise or stdf increased with ERA5 as a reference by 40% and 107%, respectively. Both stations are located in regions of steep topography where both reanalyses have significant representativeness errors compared to the GNSS observations. In the case of MKEA, the station is located at an altitude of 3729 m, whereas the altitudes of the surrounding grid points from both reanalyses are much lower. In the case of USUD, the situation is opposite, with the station is closer to the sea level than the surrounding grid points from the reanalyses.

3.1.4. Impact of the Auxiliary Data Set

The auxiliary data used in the conversion of GNSS ZTD to IWV impacts the quality of the GNSS IWV data and may lead to different segmentation results in a similar way as the processing and reference data sets. Table 2 shows that on average the mean noise and stdf are the same, but the segmentation statistics are slightly different (number of change-points, outliers, and validations). Figure 12 shows that the noise and stdf results actually change for many stations. In general, the absolute values of the noise are very close, but the relative differences are not that small. At 60% of the stations, ERAI induces larger noise than ERA5, with values up to 10–20%, while, at 40% of the stations, ERA5 yields similar or higher mean noise in ERAI, but the relative increase there is small (2.5% at maximum). These results are consistent with the representativeness differences between the reanalyses discussed above, although the pressure and temperature data are much less subject to small-scale variations than IWV.

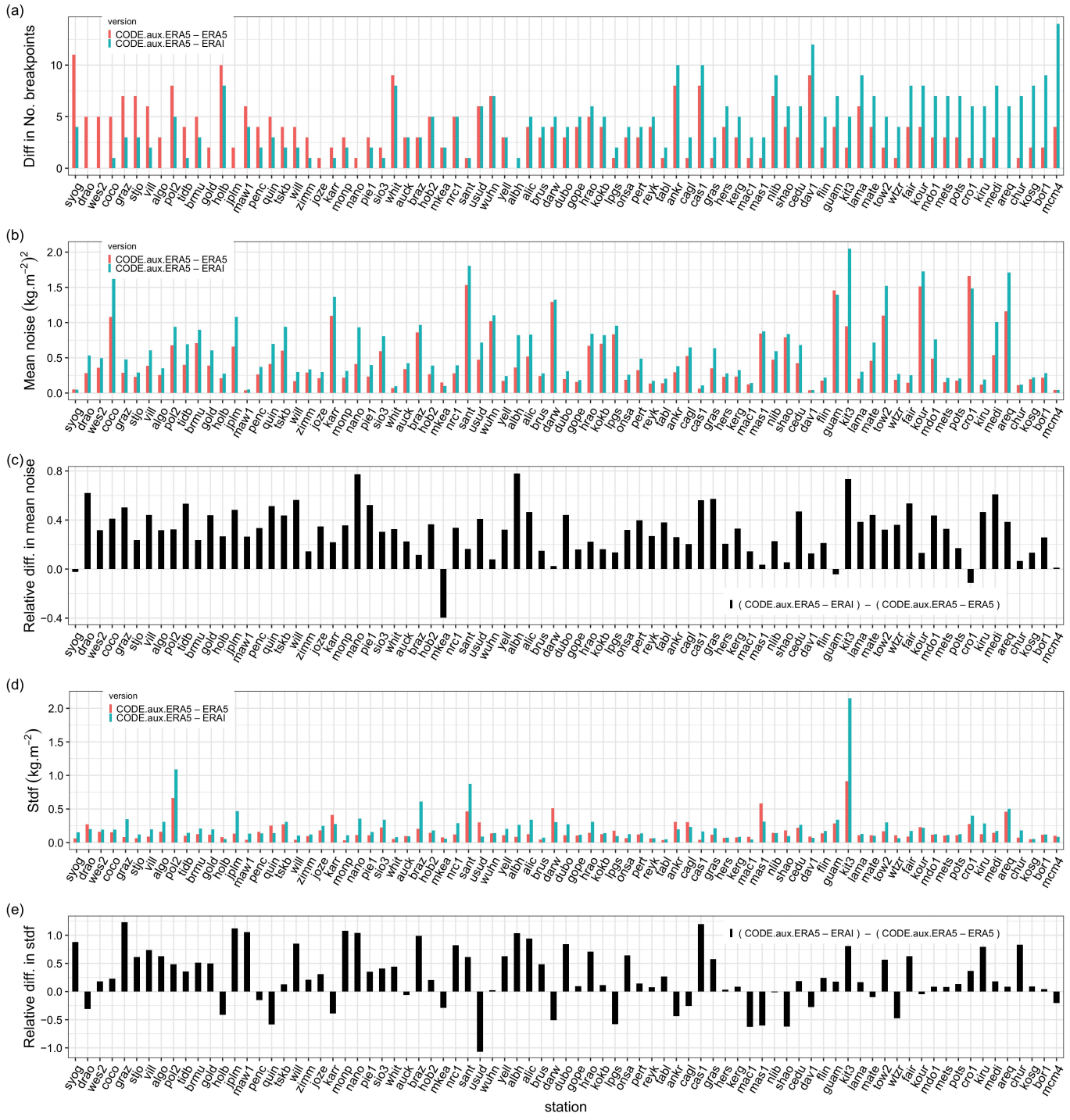


Figure 9. Similar to Figure 3 but comparing the segmentation results using two different reference data sets, ERA-Interim (ERAI) and ERA5.

The results are similar for stdf (60% of the stations have a larger periodic bias with ERAI), but the relative difference can be much higher (up to ±80%). This is because many stations are located in complex regions, such as the mountains and near the oceans. In some cases, ERA5 induces a larger periodic bias compared to the ERAI, for instance, at CHUR (Churchill, MB, Canada), KERG (Port aux Français, French Southern Territories), and TABL (Wrightwood, CA, USA).

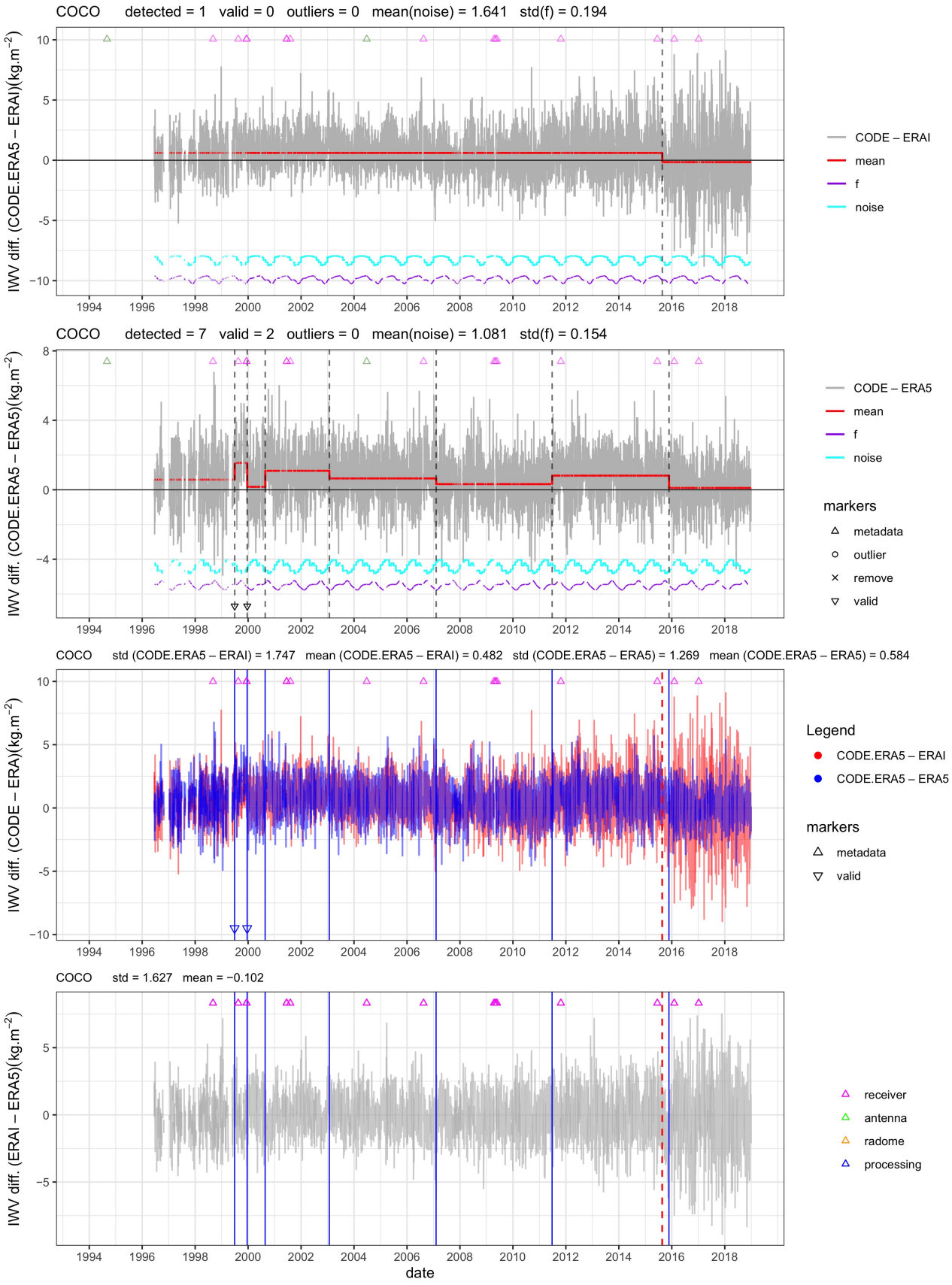


Figure 10. Similar to Figure 4, but for station COCO (Cocos (Keeling) Island, Australia).

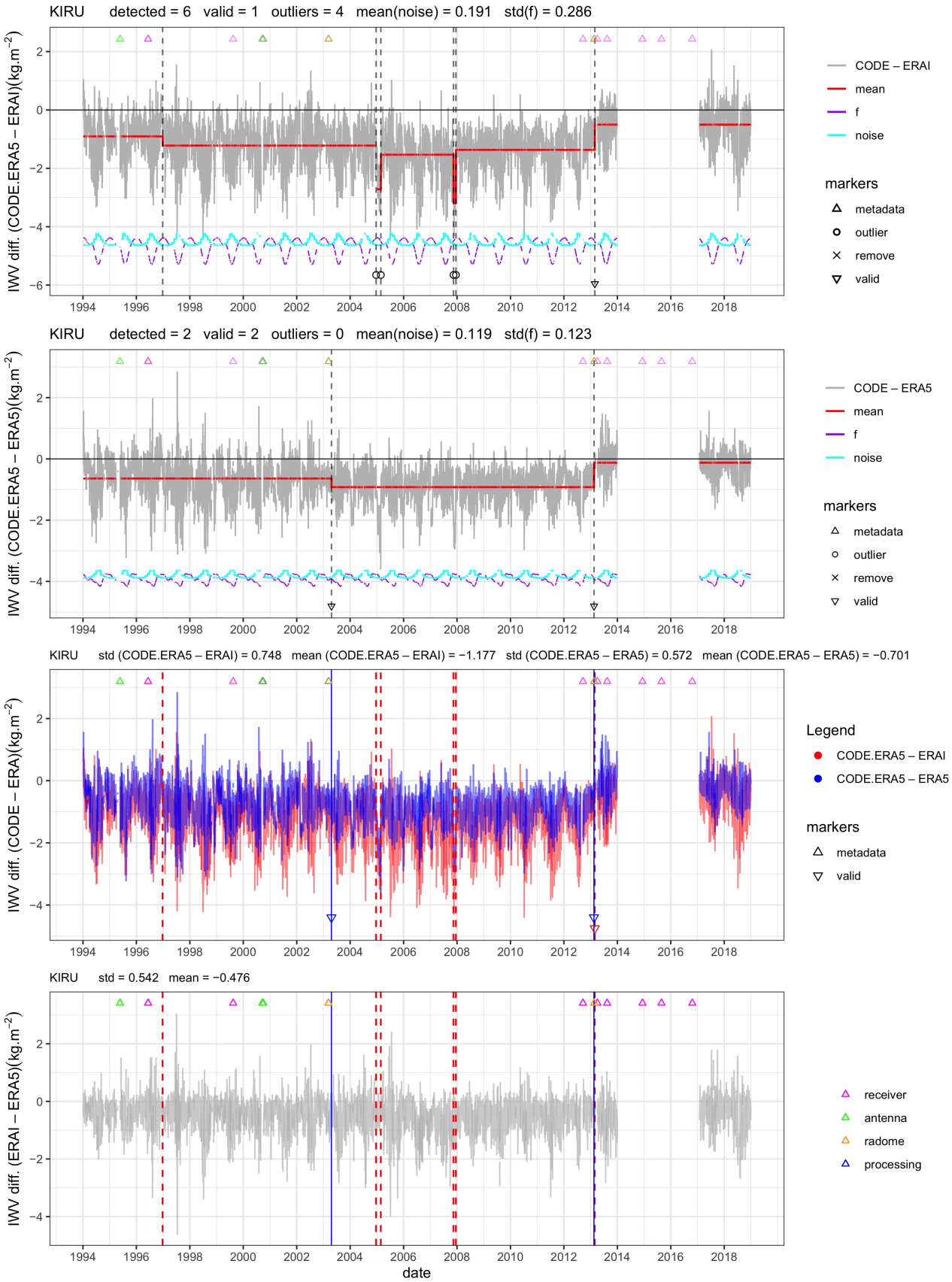


Figure 11. Similar to Figure 4, but for station KIRU (Kiruna, Sweden).

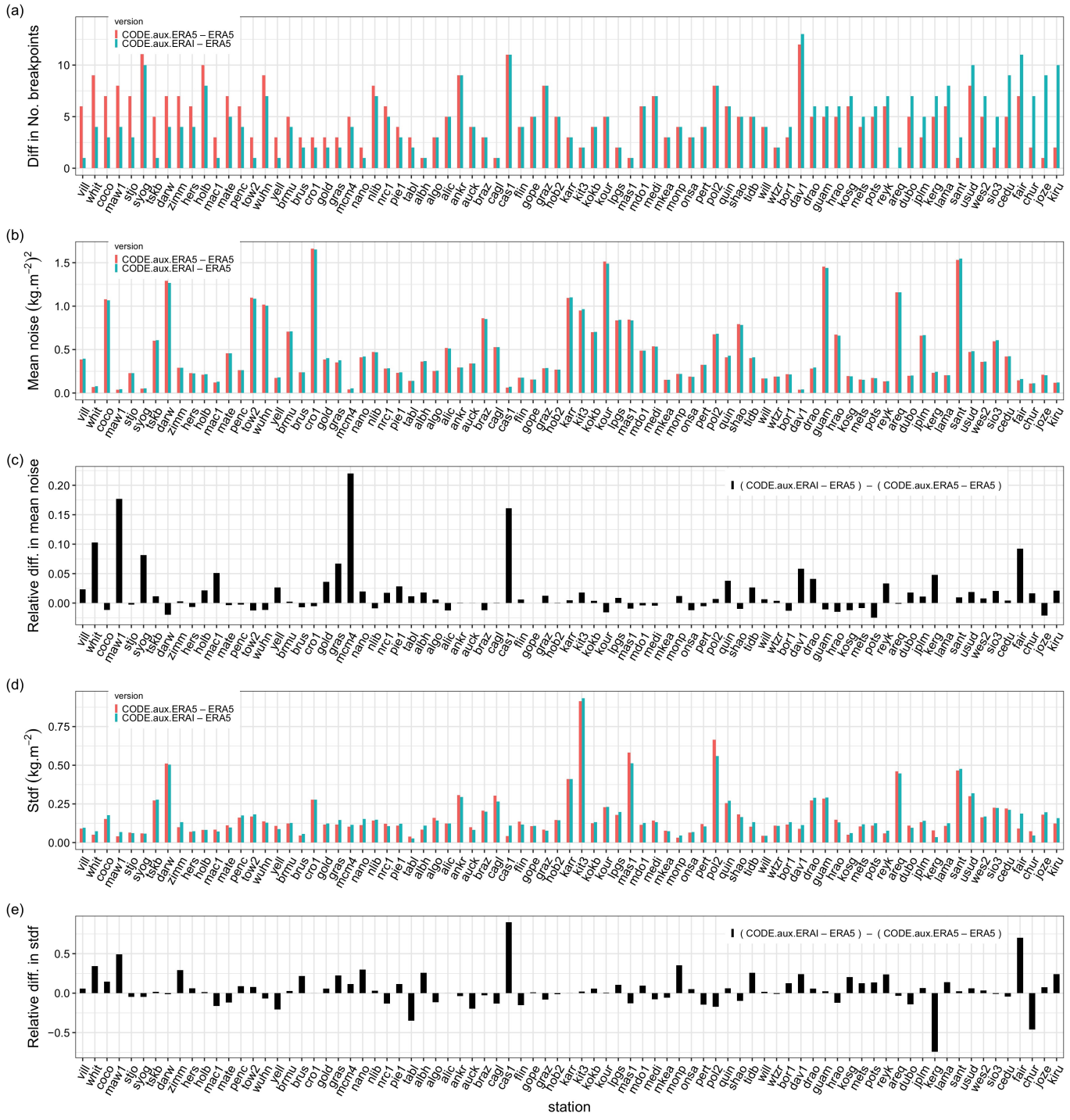


Figure 12. Similar to Figure 3, but comparing segmentation results from GNSS data sets that used two different auxiliary data, from ERA-Interim (ERAI) and ERA5.

Although the total number of change-points in the two data sets are very similar before and after screening (see Table 2), the number of change-points can be quite different from station to station (Figure 12a). The results with auxiliary data from ERAI show 18% more outliers than with ERA5. From this perspective, it is better to use ERA5. However, the percentage of similar change-points is still quite high (around 71%), which points to a moderate impact of the auxiliary data on the segmentation results in the end.

3.2. IWV Trend Estimates

3.2.1. Impact of GNSS and Reanalysis Data Set Properties on Trend Estimates

Table 3 summarizes the trend results obtained with the different GNSS data sets and the two reanalyses discussed in Section 3.1. The numbers report the mean and standard deviation of the trend estimates (in kg m^{-2} year^{-1}) over the 81 stations, as well as the number of significant trends at the 0.05 level (using a Student's *t*-test), and the standard error in the trend estimate ($1 - \sigma$). Following from Section 3.1, three-time periods, with lengths 16, 17, and 25 years, are presented, respectively.

Table 3. Summary of IWV trends from various data sets used in this work. The number of stations with significant trends at level $\alpha = 0.05$ is given in brackets. (a) GNSS data converted with auxiliary data from ERAI and segmentation applied the CODE—ERA5 IWV difference. (b) GNSS data converted with auxiliary data from ERA5 and segmentation applied the CODE—ERAI IWV difference. (c) GNSS data converted with auxiliary data from ERA5 and segmentation applied the CODE—ERA5 IWV difference.

Time Span		**1995–2010**		1994–2010		**1994–2018**	
Std error (kg m^{-2} year^{-1})		0.035		0.033		0.018	
ERAI (kg m^{-2} year^{-1})		0.018 ± 0.055 (9)		0.013 ± 0.049 (10)		0.027 ± 0.034 (37)	
ERA5 (kg m^{-2} year^{-1})		0.011± 0.052 (8)		0.008 ± 0.047 (8)		0.027±0.031 (35)	
GPS		IGS time-matched	CODE time-matched	CODE time-limited	CODE (a)	CODE (b)	CODE (c)
Raw data	IWV trend (kg m^{-2} year^{-1})	0.024 ± 0.059 (20)	0.018 ± 0.060 (18)	0.016 ± 0.060 (23)	0.033 ± 0.032 (46)	0.030 ± 0.031 (41)	0.030±0.031 (41)
	RMSE wrt ERA5 (kg m^{-2} year^{-1})	0.044	0.046	0.046	0.033	0.033	0.033
corrected IWV by validations	IWV trend (kg m^{-2} year^{-1})	0.015± 0.052 (12)	0.014± 0.052 (11)	0.011± 0.052 (15)	0.027 ± 0.027 (34)	0.025 ± 0.030 (34)	0.027 ± 0.026 (34)
	RMSE wrt ERA5 (kg m^{-2} year^{-1})	0.038	0.039	0.040	0.019	0.022	0.019
corrected IWV by all breakpoints	IWV trend (kg m^{-2} year^{-1})	0.017 ± 0.053 (9)	0.016 ± 0.054 (9)	0.012 ± 0.048 (13)	0.027 ± 0.030 (33)	0.027 ± 0.032 (35)	0.027 ± 0.030 (34)
	RMSE wrt ERA5 (kg m^{-2} year^{-1})	0.021	0.022	0.022	0.006	0.012	0.006

From the two reanalyses, we see that the mean trends are positive, indicating a net moistening, globally, with slightly different values between the three periods. This reminds us that the mean linear trends from different periods may not generally agree because they are strongly influenced by inter-annual to inter-decadal variability. However, the decrease in the standard deviation is noticeable from the shorter to the longer period (e.g., from 0.052 kg m^{-2} year^{-1} to 0.031 kg m^{-2} year^{-1} for the ERA5 data set), which indicates a decreasing influence of the inter-annual variability with time, as well as more consistent trend estimates from the global network with long time series. This decrease is also seen in the GNSS data sets, raw and corrected. It is also consistent with a decrease in the standard error with the longer time series, from 0.035 to 0.018 kg m^{-2} year^{-1}, and the subsequent increase in the number of significant trends, e.g., from 8 to 35 with ERA5. ERAI and ERA5 show different means and standard deviations in the short periods, which is not surprising according to the differences in the IWV time series extracted from the two reanalyses (see Section 3.1). The mean difference is negligible on the longer period, but the variability is slightly smaller in ERA5. The RMS difference amounts to 0.013 kg m^{-2} year^{-1} (not reported in Table 3), which indicates that there are substantial local differences in the trends from the two reanalyses. Note that the mean positive (moistening) IWV trend of 0.027 kg m^{-2} year^{-1} from the reanalyses (and also from the GNSS data) is fairly consistent with the prediction from Clausius-Clapeyron law of 7% IWV increase per 1 °C induced by the global increase in temperature of ∼1 °C over the past four decades, given a global mean IWV of 18 kg m^{-2} [50].

Next, we examine the results for the two GNSS data sets, IGS and CODE, before and after homogenization, and their differences with respect to ERA5. In this section, we consider two different corrected (homogenized) data sets. In the first one, we use only the change-points validated from the metadata, while, in the second one, we use all the detected change-points. We will refer to these data sets as "partially corrected" and "fully corrected" data sets, respectively. Ideally, we could also consider a third version where only the change-points attributed to GNSS are included, but this is not possible here because no nearby stations are available in many cases (see the discussion in Section 2). The raw GNSS trends show quite a large difference in the mean (0.024 versus 0.018 kg m^{-2} $year^{-1}$) and a RMS difference of 0.016 kg m^{-2} $year^{-1}$ (not reported Table 3). This difference is not unexpected given the differences in the data processing (Section 2) and the inhomogeneities they induce, as discussed from the segmentation results in Section 3.1. Especially, the inhomogeneities in the IGS data set due to the older antenna/radome calibration models may be a significant cause of uncertainty in the trends. The mean difference is reduced in both corrected data sets, although the RMS difference is not reduced in the partially corrected data set (0.019 kg m^{-2} $year^{-1}$), contrary to what one would expect after both data sets are homogenized. This result can be understood from the fact that the segmentation results of the IGS and CODE data sets are sometimes very different and the validated change-points may not coincide in both solutions (see Figures 4 and 6). On the other hand, the IGS and CODE GNSS fully corrected data sets are much more consistent (RMS difference of 0.006 kg m^{-2} $year^{-1}$). The latter result gives good confidence that the segmentation method is able to detect all the significant change-points in either data set. However, we know that not all these change-points may come from the GNSS time series, but some of them may be due to inhomogeneities in the reference reanalysis (in this case, ERAI). As a consequence, the fully corrected GNSS trends will be very close to the trends from the segmentation reference data set in the end. In Table 3, we give the RMS difference between the GNSS data sets and ERA5 (which is taken as another reference, although not independent from ERAI). In general, the RMS differences between the fully corrected GNSS trends and ERA5 are significantly smaller than between the raw or the partially corrected trends and the ERA5 trends. We note also that the number of significant trends is drastically reduced for both corrected GNSS data sets (from 20 to 9 for IGS and from 18 to 9 for CODE). This indicates that a large portion of uncorrected stations had significant trends because of inhomogeneities in their time series.

The results from the CODE time-limited data set are quite similar, although the change in the mean trend is reduced between the uncorrected and the corrected series, and the agreement with ERA5, in the end, is improved. The number of significant trends is also higher than with the time-matched data sets. Recall that the main difference between this data set and the CODE time-matched is only one more year and fewer gaps, but these differences can have a sensible impact on the segmentation results and trend estimates.

The results from the long time series using either ERAI or ERA5 as auxiliary data or reference data for the segmentation are presented in the rightmost part of Table 3. The mean trends from the uncorrected GNSS data are slightly larger than the trends from the reanalyses (0.030–0.033 kg m^{-2} $year^{-1}$ for GNSS compared to 0.027 kg m^{-2} $year^{-1}$ for ERAI and ERA5). The corrected GNSS series achieve closer mean trends to the reanalyses, as well as reduced RMS differences. The smallest RMS difference with respect to ERA5 is found with the fully corrected GNSS data using ERA5 as a reference, which is to be expected (ERA5 is not independent in this case). In terms of variability, the partially corrected trends show slightly smaller standard deviation, i.e., smaller spatial variability, which suggests more homogeneous and consistent trends in the global network. The slightly higher variability in the fully corrected GNSS trends might be due to some inhomogeneities in the reference series (ERA5 or ERAI). This point is further discussed in the next subsection. The impact of the segmentation reference (ERAI versus ERA5) on the corrected GNSS trends is small in terms of mean, but the RMS difference between the GNSS trends at individual stations amounts to 0.015 kg m^{-2} $year^{-1}$ for the partially corrected series

and 0.012 kg m^{-2} year^{-1} for the fully corrected series (not shown). The impact of the auxiliary data is significantly smaller, with a RMS difference between the trends using ERAI and ERA5 as auxiliary of 0.008 kg m^{-2} year^{-1} for the partially corrected series and 0.002 kg m^{-2} year^{-1} for the fully corrected series (not shown).

3.2.2. Impact of Homogenization on GNSS Trend Estimates

In this sub-section, we analyze in more details our 'best' data set at hand, i.e., the long CODE GNSS series (1994–2018) using ERA5 as auxiliary data and reference for the segmentation. Figure 13a shows the IWV trend estimates for the GNSS data, uncorrected and corrected, and ERA5. A majority of the GNSS stations have positive trends (89% with the partially corrected data, 86% with the fully corrected data, versus 80% for the ERA5 data), consistent with the overall positive mean trend of 0.027 kg m^{-2} year^{-1} reported in Table 3. Among these, only $\sim$ 41% of the corrected GNSS trends are significant at $p = 0.05$ (or t = 1.99, see Figure 13b), versus 49% with the uncorrected GNSS data, and 41% with ERA5. The largest positive trend is found for KOUR (Kourou, French Guiana), reaching a value of $\sim$0.110 kg m^{-2} year^{-1} (t = 5) for the uncorrected and partially corrected trends, and 0.150–0.153 kg m^{-2} year^{-1} (t $\sim$ 7) for the fully corrected GNSS data and ERA5. This station, as well as a few others in the Tropics (KOKB, GUAM, BRMU, CRO1, etc.), shows consistent and significant moistening over the past 2.5 decades. Strong moistening is also found at several Mediterranean stations (MEDI, CAGL, MATE), confirming the strong warming in this area [51]. A few stations show consistent drying from GNSS data (uncorrected and corrected) and ERA5, mostly in arid regions, such as JPLM (Pasadena, CA, USA), HRAO (Krugersdorp, South Africa), ALIC (Alice Springs, Australia), and SANT (Santiago, Chile).

A striking feature in Figure 13 is that many uncorrected GNSS trends are quite large, and significant, while the corrected trends are much smaller, and are, in some cases, insignificant. The trends of the partially corrected series are significantly different from the uncorrected trend (with a RMS difference of 0.022 kg m^{-2} year^{-1}), although they agree with each other better than with ERA5 (0.033 kg m^{-2} year^{-1}). On the other hand, the trends of the fully corrected GNSS data are much closer to ERA5 (with a RMS difference of 0.006 kg m^{-2} year^{-1}) than to the uncorrected trends (RMS difference of 0.031 kg m^{-2} year^{-1}). The RMS difference between the partially corrected and the fully corrected trends is 0.016 kg m^{-2} year^{-1}. It is actually expected that the corrected trends will tend to align with the reference data (in this case, ERA5). However, the partially corrected trends are relatively independent from the reference since only about 36.5% of the detected change-points are validated. It is also noticeable that the partial correction has a strong impact on the trends at many sites, such as a change in the sign (e.g., GOPE, KARR, VILL, etc.), a change from significant to insignificant (e.g., HERS, SHAO, SYOG, WUHN, etc.), or vice versa (GOPE, KOKB, TIDB, etc.). The trends of the fully corrected data get closer to ERA5.

Four cases will be discussed in more detail in the following, where the IWV trends change from significant to insignificant, or vice versa. The corresponding time series and change-points are shown in Figure 14.

Firstly, we examine the case of station WUHN (Wuhan, China), where the uncorrected GNSS trend is positive and significant (0.115 kg m^{-2} year^{-1}, t = 2.44), while the fully corrected trend and the ERA5 trend are of opposite sign ($-$0.031 and $-$0.037 kg m^{-2} year^{-1}, respectively) but not significant (t = $-$0.65 and $-$0.79). At this site, the change in sign of the trend is suspected to arrive from a strong downward shift in the ERA5 time series, which was already commented by Parracho et al. [14] and attributed to a change in the radiosonde type in Wuhan in 2006 impacting the ERA-Interim reanalysis. Parracho et al. [14] also noticed an extended region of a negative trend in eastern China, in contradiction with the GNSS observations and the MERRA-2 reanalysis. A similar shift is actually detected in the nearby station SHAO (Sheshan, China), where the uncorrected GNSS trend of 0.086 kg m^{-2} year^{-1} is decreased to 0.043 kg m^{-2} year^{-1} in the fully corrected data. The

ERA5 trend is even smaller (0.037 kg m^{-2} year^{-1}), although not negative. At both WUHN and SHAO, the GNSS trend changes from significant to insignificant after correction.

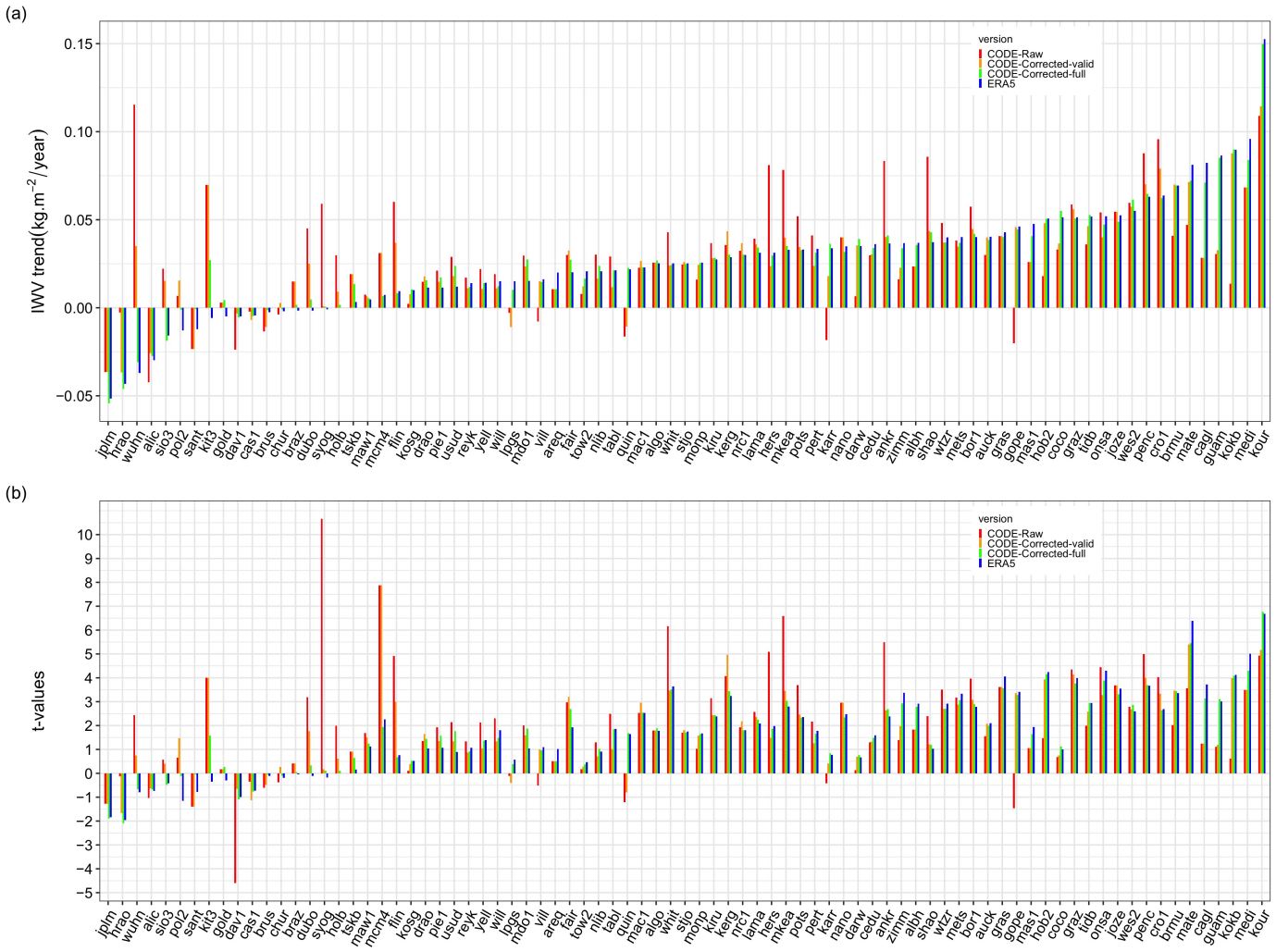


Figure 13. Comparison of: (**a**) trend estimates kg m^{-2} year^{-1}, and (**b**) t-values from different uncorrected (raw) and corrected GNSS data (CODE with auxiliary ERA5 and reference ERA5) and ERA5 at 81 common stations.

Secondly, let us examine the case of station HERS (Hailsham, UK). The segmentation detected six change-points, among which only two are validated with the metadata. Overall, the mean shifts are going upwards, meaning that the inhomogeneities induce a spurious positive trend in the GNSS series. The correction of the GNSS series for two validated change-points has a strong impact on the trend, decreasing it from 0.081 to 0.024 kg m^{-2} year^{-1} and from significant to insignificant (t-value from 5.1 to 1.5). Including the 4 additional change-points has a further, although small, effect, leading to a final GNSS trend of 0.030 kg m^{-2} year^{-1}, close to the ERA5 trend (0.031 kg m^{-2} year^{-1}). Two nearby stations (HERT and HRM1) could be used in the attribution step to confirm the two validated change-points but not the other ones. The other change-points could not be tested. The impact of the correction is substantial and seems justified at this station, with a final trend reduced by -0.051 kg m^{-2} year^{-1}, i.e., a factor of 2.7.

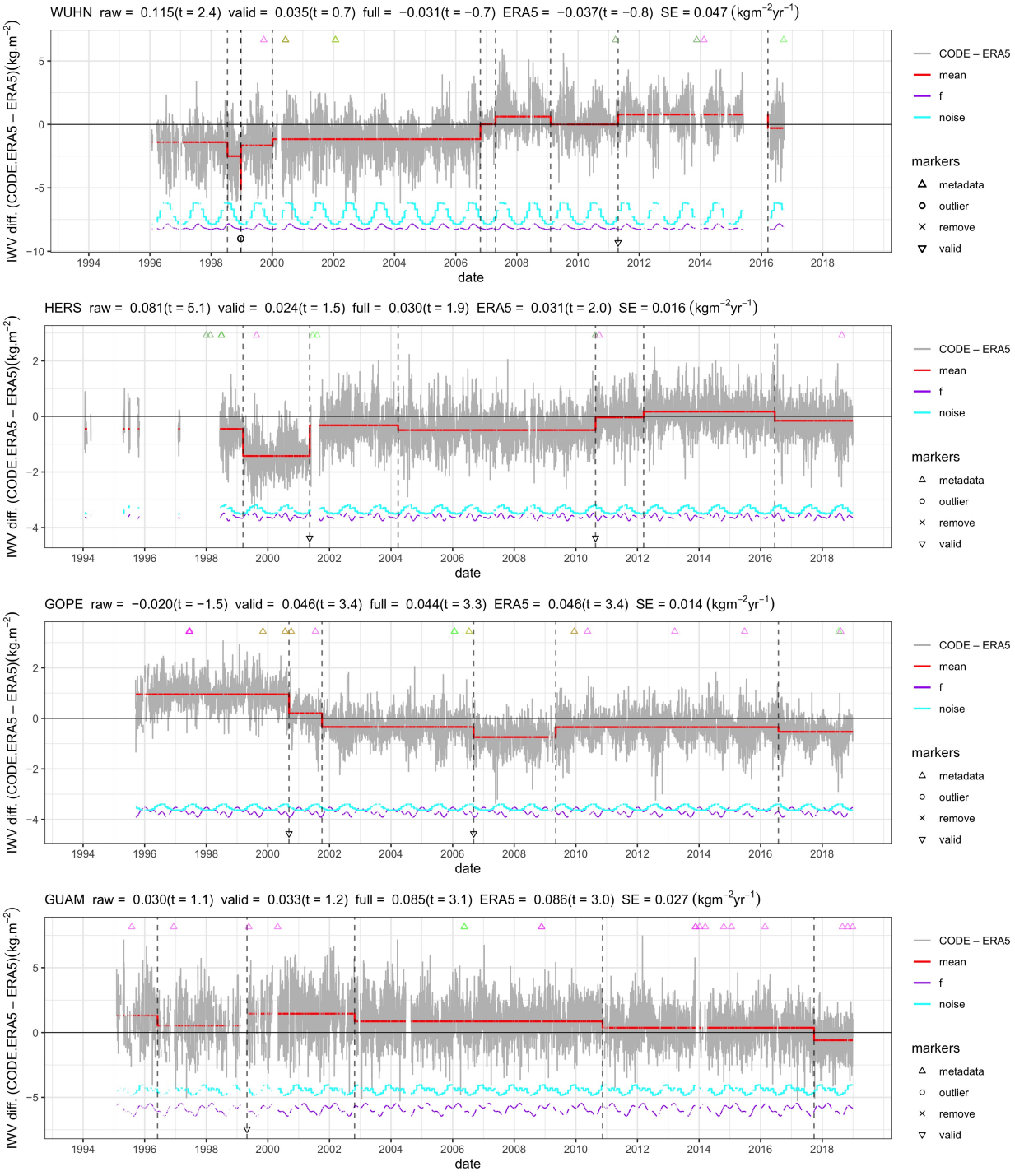


Figure 14. Similar to the upper plot in Figure 4, but for different stations: HERS (Hailsham, United Kingdom), GOPE (Ondrejov, Czech Republic), KOKB (Waimea, United States), and GUAM (Dededo, Guam). The IWV differences are computed as GNSS - ERA5, where GNSS is converted using auxiliary data from ERA5, and the segmentation is run with ERA5 as a reference.

Next, we examine station GOPE (Ondrejov, Czech Republic), which has a strong significant trend after correction but insignificant before. GOPE is a special case, which has a negative trend in the raw data in contradiction was many surrounding stations in Europe, such as ZIMM (Switzerland), WTZR (Germany), and BOR1 (Poland), which have positive trends. This feature was already noticed by Parracho et al. [14] from the uncorrected IGS data set over the shorter period (1995–2010). Figure 14 shows that the mean shifts are going

downwards, so inducing a negative trend in the GNSS series compared to ERA5. Two change-points are validated with the metadata. After correction of these change-points, the trend goes from a insignificant drying of -0.020 kg m^{-2} year^{-1} to a significant moistening of 0.046 kg m^{-2} year^{-1}. Three other change-points have a minor impact (the fully corrected trend is 0.044 kg m^{-2} year^{-1}) because the most important break in 2000 is validated. For this station, we could also test the attribution with several nearby stations collocated with station WTZR (distant by 162 km). The two validated change-points, as well as the one in 2001, could be attributed to GOPE.

The last example is station GUAM (Dededo, Guam), in the western tropical Pacific, which has a similar large trend in ERA5 to KOKB, another station in the Pacific Ocean. The trends are very different between the partially and fully corrected GNSS series at GUAM because only one change-point is validated, and it is located near the beginning of the series. The last three change-points have a strong impact on the GNSS correction, although their origin is questionable. Indeed, they are located quite far away from any known equipment change reported in the metadata. The last change-point (on 26 September 2017) could be checked in the attribution step with the nearby station GUUG (Mangilao, USA), located at a distance of 18 km from GUAM. Comparing the GNSS series at GUUG to the ERA5 series at GUAM revealed a significant change in mean on this date. From this result, we should attribute this change-point to the ERA5 series and not the GNSS series. At this site, thus, we also suspect the other unvalidated change-points to be due to ERA5. This assumption may be further checked by inspecting observation statistics from the assimilation system, but this is left for future work.

4. Conclusions

This study investigated the sensitivity of the segmentation method and the IWV trend estimates to different GNSS and reference data properties. It was shown that the GNSS processing methods and the reference data (ERAI versus ERA5) have the strongest impact on the segmentation results (i.e., number and positions of change-points), while the impact is weaker when the length of the time series (17 versus 25 years) or the auxiliary data (ERAI versus ERA5) are changed. Changing the latter two was shown to achieve 81% and 71% similar change-points, but only 45–49% when the GNSS data set or the reference was changed. These discrepancies in the results indicate that the segmentation is sensitive to small changes in the mean signal, in the noise, and in the periodic bias. These features are determined by several aspects of the GNSS processing procedure, especially: the a priori ZHD correction, the antenna/radome calibration model, the mapping functions, and the elevation cutoff angle. The more recent and more accurate ZHD correction, antenna/radome calibration model, and mapping functions used in the CODE reprocessing achieve smaller noise and periodic bias but, at the same time, lead to an increase in the number of detected change-points because the segmentation is able to detect smaller changes in the mean. The same is observed when the reference is changed from ERAI to ERA5. The reduction in the noise and periodic bias in ERA5 is partly due to the better spatial resolution and partly to the more recent atmospheric model and assimilation of more observations. In terms of the percentage of validation of the detected change-points with respect to the GNSS metadata, the length of the time series has the strongest impact. The average percentage is $\sim$30% with the short time series (16 or 17 years), while it rises to 34–36% with the long time series (25 years), for a window of $\pm$62 days. This difference might be due to the particular penalty function that we used here [46,47]. The penalty function might be calibrated differently to achieve a better balance between the probability of detection and the probability of false detection to our particular data [6].

This study also points to the strong impact of the inhomogeneity correction on the estimated trends. When only the change-points validated with the metadata are used, the impact of the correction is substantial on the global mean trend, on the dispersion, and on the RMS difference with respect to ERA5. When all the detected change-points are used, the mean and dispersion are again modified, but to a lesser extent, and the RMS

difference with ERA5 is further reduced. The latter feature is expected when ERA5 or ERAI are used as a reference in the segmentation, especially since a few change-points may be due to the reanalyses (e.g., suspected at stations WUHN, GUAM). The uncertainty in the estimated trends from the use of the different reprocessed GNSS data sets or reference reanalysis in the segmentation is about 0.015–0.019 kg m^{-2} $year^{-1}$ (RMS difference between the tested pairs of data sets) when only the validated change-points are used and 0.002 to 0.012 kg m^{-2} $year^{-1}$ when all the change-points are used. The auxiliary data has a marginal impact on the trends. The longer time series (25 years) provides higher accuracy in the trend estimates, with a mean standard error of 0.018 kg m^{-2} $year^{-1}$ and a dispersion of ∼0.03 kg m^{-2} $year^{-1}$ throughout the global network. The homogenized GNSS trends and both reanalyses agree in the global mean IWV trend estimate of 0.027 kg m^{-2} $year^{-1}$, which indicates a global moistening over the past 2.5 decades at a rate close to the 7% per degree of warming predicted by Clausius Clapeyron equation [50].

The main limitation of our current homogenization method is the attribution step (Figure 2), which could not be used to check all the change-points because of the lack of nearby stations. In future work, we plan to include reprocessed GNSS data from additional stations in regions where dense networks are available (e.g., Europe, USA, Japan, etc.). Some limitations of the current segmentation method were highlighted, as well, especially the dependence on the length of the time series, which may be mitigated by careful calibration of the penalty function. The segmentation method is also sensitive to the long-term variations in the noise and bias, which are not specifically modeled here (we assume both are periodic with a fundamental period of 1 year). At some sites, strong inter-annual or decadal variations in the noise and/or bias were captured by the segmentation. These spurious change-points need to the detected and removed, unless the long-term variations are reduced, e.g., from the use of a better reference data set, such as a nearby GNSS station, rather than a reanalysis.

Finally, this study also helped to better understand the performance and limitations of the GNSS data processing procedures. This knowledge can be helpful in the future to reprocess the GNSS data and produce better "homogenized" daily IWV time series at the processing level.

Author Contributions: E.L., O.B. and K.N.N. developed the statistical conceptualization methods of the work. Data and climatological aspects were dealt with by O.B., and computational issues were handled by O.B. and K.N.N., all shared in the writing of the manuscript. A.Q. developed the segmentation method and R package used in this study. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the CNRS program LEFE/INSU through the project VEGAN.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The daily IWV data set based on the IGS repro1 ZTD data (120 stations, period 1995–2010) is available from https://doi.org/10.14768/06337394-73a9-407c-9997-0e380dac5591, accessed on 30 July 2021 [43]. The daily IWV data set based on the CODE REPRO2015 ZTD data (436 stations, period 1994–2018) is available from https://dx.doi.org/10.25326/18, accessed on 30 July 2021 [44].

Acknowledgments: The authors are grateful to AERIS, the French data and service center for atmosphere, for providing the ERA-Interim and ERA5 reanalysis data, and hosting the GNSS IWV data sets. The contribution of the fourth author has been conducted as part of the project Labex MME-DII (ANR11-LBX-0023- 01) and within the FP2M federation (CNRS FR 2036).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CODE	Center for Orbit Determination in Europe
ECMWF	European Center for Medium-Range Weather Forecasts
ERP	Earth Rotation Parameter
ERAI	ECMWF reanalysis Interim
ERA5	ECMWF reanalysis v5
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GMF	Global Mapping Function
IGS	International GNSS Service
IWV	Integrated water vapor
JPL/NASA	NASA's Jet Propulsion Laboratory
NMF	Neill Mapping Function
PC0	Phase Center Offset
PCV	Phase Center Variation
PWV	Precipitable Water Vapor
VMF	Vienna Mapping Function
ZHD	Zenith Hydrostatic Delay
ZTD	Zenith Tropospheric Delay
ZWD	Zenith Wet Delay

References

1. Caussinus, H.; Mestre, O. Detection and correction of artificial shifts in climate series. *J. R. Stat. Soc. Ser. C Appl. Stat.* **2004**, *53*, 405–425. [CrossRef]
2. Reeves, J.; Chen, J.; Wang, X.L.; Lund, R.; Lu, Q.Q. A Review and Comparison of Changepoint Detection Techniques for Climate Data. *J. Appl. Meteorol. Climatol.* **2007**, *46*, 900–915. [CrossRef]
3. Jones, P.D.; Raper, S.C.B.; Bradley, R.S.; Diaz, H.F.; Kellyo, P.M.; Wigley, T.M.L. Northern Hemisphere Surface Air Temperature Variations: 1851–1984. *J. Clim. Appl. Meteorol.* **1986**, *25*, 161–179. [CrossRef]
4. Easterling, D.; Peterson, T. A new method for detecting undocumented discontinuities in climatological time series. *Int. J. Climatol.* **1995**, *15*, 369–377. [CrossRef]
5. Peterson, T.C.; Easterling, D.R.; Karl, T.R.; Groisman, P.; Nicholls, N.; Plummer, N.; Torok, S.; Auer, I.; Boehm, R.; Gullett, D.; et al. Homogeneity adjustments of in situ atmospheric climate data: A review. *Int. J. Climatol. A J. R. Meteorol. Soc.* **1998**, *18*, 1493–1517. [CrossRef]
6. Menne, M.J.; Williams, C.N. Detection of Undocumented Changepoints Using Multiple Test Statistics and Composite Reference Series. *J. Clim.* **2005**, *18*, 4271–4286. [CrossRef]
7. Szentimrey, T. Development of MASH homogenization procedure for daily data. In Proceedings of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases, WCDMP-No. 71, Budapest, Hungary, 29 May–2 June 2008; pp. 123–130.
8. Costa, A.C.; Soares, A. Homogenization of Climate Data: Review and New Perspectives Using Geostatistics. *Math. Geosci.* **2009**, *41*, 291–305. [CrossRef]
9. Venema, V.K.C.; Mestre, O.; Aguilar, E.; Auer, I.; Guijarro, J.A.; Domonkos, P.; Vertacnik, G.; Szentimrey, T.; Stepanek, P.; Zahradnicek, P.; et al. Benchmarking homogenization algorithms for monthly data. *Clim. Past* **2012**, *8*, 89–115. [CrossRef]
10. Bevis, M.; Bussinger, S.; Herring, T.A.; Rocken, C.; Anthes, R.A.; Ware, R.H. GPS Meteorology: Remote Sensing of Atmospheric Water Vapor Using the Global Positioning System. *J. Geophys. Res.* **1992**, *97*, 15787–15801. [CrossRef]
11. Bock, O.; Bosser, P.; Bourcy, T.; David, L.; Goutail, F.; Hoareau, C.; Keckhut, P.; Legain, D.; Pazmino, A.; Pelon, J.; et al. Accuracy assessment of water vapour measurements from in situ and remote sensing techniques during the DEMEVAP 2011 campaign at OHP. *Atmos. Meas. Tech.* **2013**, *6*, 2777–2802. [CrossRef]
12. Van Malderen, R.; Brenot, H.; Pottiaux, E.; Beirle, S.; Hermans, C.; De Maziere, M.; Wagner, T.; De Backer, H.; Bruyninx, C. A multi-site intercomparison of integrated water vapour observations for climate change analysis. *Atmos. Meas. Tech.* **2014**, *7*, 2487–2512. [CrossRef]
13. Ning, T.; Wickert, J.; Deng, Z.; Heise, S.; Dick, G.; Vey, S.; Schöne, T. Homogenized Time Series of the Atmospheric Water Vapor Content Obtained from the GNSS Reprocessed Data. *J. Clim.* **2016**, *29*, 2443–2456. [CrossRef]
14. Parracho, A.C.; Bock, O.; Bastin, S. Global IWV trends and variability in atmospheric reanalyses and GPS observations. *Atmos. Chem. Phys.* **2018**, *18*, 16213–16237. [CrossRef]
15. Dee, D.P.; Uppala, S.M.; Simmons, A.J.; Berrisford, P.; Poli, P.; Kobayashi, S.; Andrae, U.; Balmaseda, M.A.; Balsamo, G.; Bauer, P.; et al. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **2011**, *137*, 553–597. 10.1002/qj.828. [CrossRef]

16. Gelaro, R.; McCarty, W.; Suárez, M.J.; Todling, R.; Molod, A.; Takacs, L.; Randles, C.A.; Darmenov, A.; Bosilovich, M.G.; Reichle, R.; et al. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Clim.* **2017**, *30*, 5419–5454. [CrossRef] [PubMed]
17. Bock, O.; Parracho, A. Consistency and representativeness of integrated water vapour from ground-based GPS observations and ERA-Interim reanalysis. *Atmos. Chem. Phys.* **2019**, *19*, 9453–9468. [CrossRef]
18. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. The ERA5 Global Reanalysis. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 1999–2046. [CrossRef]
19. Vey, S.; Dietrich, R.; Fritsche, M.; Rülke, A.; Steigenberger, P.; Rothacher, M. On the homogeneity and interpretation of precipitable water time series derived from global GPS observations. *J. Geophys. Res. Atmos.* **2009**, *114*. [CrossRef]
20. Domonkos, P. Adapted Caussinus-Mestre Algorithm for Networks of Temperature series (ACMANT). *Int. J. Geosci.* **2011**, *2*, 293–309. [CrossRef]
21. Quarello, A. Development of New Homogenisation Methods for GNSS Atmospheric Data. Application to the Analysis of Climate Trends and Variability. Ph.D. Thesis, Sorbonne Universite, Paris, France, 2020.
22. Van Malderen, R.; Pottiaux, E.; Klos, A.; Domonkos, P.; Elias, M.; Ning, T.; Bock, O.; Guijarro, J.; Alshawaf, F.; Hoseini, M.; et al. Homogenizing GPS Integrated Water Vapor Time Series: Benchmarking Break Detection Methods on Synthetic Data Sets. *Earth Space Sci.* **2020**, *7*, e2020EA001121. [CrossRef]
23. Teunissen, P.J.; Montenbruck, O. (Eds.) *Springer Handbook of Global Navigation Satellite Systems*; Springer International Publishing: Berlin/Heidelberg, Germany, 2017. [CrossRef]
24. Petit, G.; Luzum, B. *IERS 2010 Conventions*; Technical Report; IERS: Frankfurt-am-Main, Germany, 2010.
25. Guerova, G.; Jones, J.; Douša, J.; Dick, G.; de Haan, S.; Pottiaux, E.; Bock, O.; Pacione, R.; Elgered, G.; Vedel, H.; et al. Review of the state of the art and future prospects of the ground-based GNSS meteorology in Europe. *Atmos. Meas. Tech.* **2016**, *9*, 5385–5406. [CrossRef]
26. Davis, J.L.; Herring, T.A.; Shapiro, I.I.; Rogers, A.E.E.; Elgered, G. Geodesy by radio interferometry: Effects of atmospheric modeling errors on estimates of baseline length. *Radio Sci.* **1985**, *20*, 1593–1607. [CrossRef]
27. Bock, O. Standardization of ZTD screening and IWV conversion. In *Advanced GNSS Tropospheric Products for Monitoring Severe Weather Events and Climate: COST Action ES1206 Final Action Dissemination Report*; Jones, J., Guerova, G., Douša, J., Dick, G., de Haan, S., Pottiaux, E., Bock, O., Pacione, R., van Malderen, R., Eds.; Springer International Publishing AG: Cham, Switzerland. 2020; Chapter 5, pp. 314–324. [CrossRef]
28. Bock, O.; Bosser, P.; Flamant, C.; Doerflinger, E.; Jansen, F.; Fages, R.; Bony, S.; Schnitt, S. Integrated water vapour observations in the Caribbean arc from a network of ground-based GNSS receivers during EUREC^{4}A. *Earth Syst. Sci. Data* **2021**, *13*, 2407–2436. [CrossRef]
29. Tregoning, P.; Herring, T.A. Impact of a priori zenith hydrostatic delay errors on GPS estimates of station heights and zenith total delays. *Geophys. Res. Lett.* **2006**, *33*. 10.1029/2006GL027706. [CrossRef]
30. Schmid, R.; Steigenberger, P.; Gendt, G.; Ge, M.; Rothacher, M. Generation of a consistent absolute phase-center correction model for GPS receiver and satellite antennas. *J. Geod.* **2007**, *81*, 781–798. [CrossRef]
31. Schmid, R.; Dach, R.; Collilieux, X.; Jäggi, A.; Schmitz, M.; Dilssner, F. Absolute IGS antenna phase center model igs08.atx: Status and potential improvements. *J. Geod.* **2015**, *90*, 343–364. [CrossRef]
32. Ning, T.; Wang, J.; Elgered, G.; Dick, G.; Wickert, J.; Bradke, M.; Sommer, M.; Querel, R.; Smale, D. The uncertainty of the atmospheric integrated water vapour estimated from GNSS observations. *Atmos. Meas. Tech.* **2016**, *9*, 79–92. [CrossRef]
33. Zumberge, J.F.; Heflin, M.B.; Jefferson, D.C.; Watkins, M.M. Precise point positioning for the efficient and robust analysis of GPS data from large networks. *J. Geophys. Res.* **1997**, *102*, 5005–5017. doi:10.1029/96JB03860. [CrossRef]
34. Byun, S.H.; Bar-Sever, Y.E. A new type of troposphere zenith path delay product of the international GNSS service. *J. Geod.* **2009**, *83*, 1–7. [CrossRef]
35. Boehm, J.; Niell, A.E.; Tregoning, P.; Schuh, H. The Global Mapping Function (GMF) : A new empirical mapping function based on numerical weather model data. *Geophys. Res. Lett.* **2006**, *33*, L07304. [CrossRef]
36. Niell, A.E. Global mapping functions for the atmosphere delay at radio wavelengths. *J. Geophys. Res. Solid Earth* **1996**, *101*, 3227–3246. [CrossRef]
37. Susnik, A.; Dach, R.; Villiger, A.; Maier, A.; Arnold, D.; Schaer, S.; Jäggi, A. CODE Reprocessing Product Series. 2016. Available online: https://boris.unibe.ch/80011/ (accessed on 25 August 2021). [CrossRef]
38. Dach, R.; Lutz, S.; Walser, P.; Fridez, P. Bernese GNSS Software Version 5.2; User Manual. 2015. Available online: https://boris.unibe.ch/72297/ (accessed on 8 May 2012). [CrossRef]
39. Dach, R.; Schaer, S.; Arnold, D.; Orliac, E.; Prange, L.; Susnik, A.; Villiger, A.; Jäggi, A. *CODE Final Product Series for the IGS*; 2018. Available online: https://boris.unibe.ch/119490/ (accessed on 25 August 2021). [CrossRef]
40. Boehm, J.; Werl, B.; Schuh, H. Troposphere mapping functions for GPS and very long baseline interferometry from European Centre for Medium-Range Weather Forecasts operational analysis data. *J. Geophys. Res.* **2006**, *11*, 2406. [CrossRef]
41. Bock, O.; Pacione, R.; Ahmed, F.; Araszkiewicz, A.; Bałdysz, Z.; Balidakis, K.; Barroso, C.; Bastin, S.; Beirle, S.; Berckmans, J.; et al. Use of GNSS Tropospheric Products for Climate Monitoring (Working Group 3). In *Advanced GNSS Tropospheric Products for Monitoring Severe Weather Events and Climate*; Jones, J., Guerova, G., Douša, J., Dick, G., de Haan, S., Pottiaux, E., Bock, O., Pacione, R., van Malderen, R., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 267–402

42. Stepniak, K.; Bock, O.; Wielgosz, P. Reduction of ZTD outliers through improved GNSS data processing and screening strategies. *Atmos. Meas. Tech.* **2018**, *11*, 1347–1361. [CrossRef]
43. Bock, O. GPS Data: Daily and Monthly Reprocessed IWV Data from 120 Global GPS Stations, Version 1.2. 2016. Available online: https://observations.ipsl.fr/espri/metadata/global_gps_iwv_v1.2.html/ (accessed on 25 August 2021). [CrossRef]
44. Bock, O. Global GNSS IWV Data at 436 Stations over the 1994–2018 Period. 2019. Available online: https://www.aeris-data.fr/metadata/metadata/?5829bd3d-4593-4e66-bce2-c8c6311360af (accessed on 25 August 2012). [CrossRef]
45. Quarello, A.; Bock, O.; Lebarbier, E. A new segmentation method for the homogenisation of GNSS-derived IWV time-series. *arXiv* **2020**, arXiv:2005.04683.
46. Birgé, L.; Massart, P. Gaussian model selection. *J. Eur. Math. Soc.* **2001**, *3*, 203–268. [CrossRef]
47. Lebarbier, E. Detecting Multiple Change-Points in the Mean of Gaussian Process by Model Selection. *Signal Process.* **2005**, *85*, 717–736. [CrossRef]
48. Weatherhead, E.C.; Reinsel, G.C.; Tiao, G.C.; Meng, X.; Choi, D.; Cheang, W.; Keller, T.; DeLuisi, J.; Wuebbles, D.J.; Kerr, J.B.; et al. Factors affecting the detection of trends: Statistical considerations and applications to environmental data. *J. Geophys. Res. Atmos.* **1998**, *103*, 17149–17161. [CrossRef]
49. Koulali, A.; Clarke, P.J. Effect of antenna snow intrusion on vertical GPS position time series in Antarctica. *J. Geod.* **2020**, *94*. [CrossRef]
50. Dunn, R.J.H.; Stanitski, D.M.; Gobron, N.; Willett, K.M. Global Climate [in "State of the Climate in 2019"]. *Bull. Am. Meteorol. Soc.* **2020**, *101*, S9–S127. [CrossRef]
51. Ulbrich, U.; Lionello, P.; Belušić, D.; Jacobeit, J.; Knippertz, P.; Kuglitsch, F.G.; Leckebusch, G.C.; Luterbacher, J.; Maugeri, M.; Maheras, P.; et al. 5—Climate of the Mediterranean: Synoptic Patterns, Temperature, Precipitation, Winds, and Their Extremes. In *The Climate of the Mediterranean Region*; Lionello, P., Ed.; Elsevier: Oxford, UK, 2012; pp. 301–346.

Chapter 4

Development of the attribution method

4.1 Paper No. 2: A statistical method for the attribution of change-points in segmented IWV difference time series

This chapter comprises two main sections. The first one corresponds to a paper, submitted to the International Journal of Climatology (presently under review) which describes the attribution method developed in this thesis and the results on a real data set of 81 main stations. The second section presents additional research conducted alongside the paper's topic. It includes namely a comprehensive assessment of the noise model identification and parameter estimation tools used in the paper, based on numerical simulations. It also explores some alternative methodologies that may help to further increase the efficiency of several parts (regression, prediction) of the attribution method.

4.1.1 Abstract

Many segmentation or change-point detection methods used for the homogenization of climate time series from station data use a reference series against which the station data is compared. The main advantage of this approach is to remove the common climate signal and thus improve the detection power. One drawback is that it is difficult to decide whether the detected change-point is due to the main series or to the reference. A so-called attribution procedure is typically applied in a post-processing step for each detected change-point and each main station. This paper describes a new statistical method for the attribution of detected change-points in GNSS minus reanalysis series of Integrated Water Vapour (IWV). It works by combining the GNSS and reanalysis data from the main station with similar data from one or several nearby stations. The paired data from one main station and one nearby station form a set of four base series. The six series of differences formed from these four base series are tested for a significant jump at the time of the detected change-point in the main station. A statistical predictive rule is used to attribute the change-points in the four base series from the six test results. Original aspects of our method are: 1) the significance test, which is based on a generalized linear regression approach, taking both heteroscedasticity and autocorrelation into account; 2) the predictive rule, which uses a machine learning method and is constructed from the test results obtained with the real data by using a resampling strategy. Four popular machine learning methods have been compared using cross-validation and the best one was applied to a real data set (49 main stations with 114 change-points). The results depend on the choice of the test significance level and the aggregation method combining the prediction results when

several nearby stations are available. We find that 62% of the change-points are attributed to GNSS, 19% to the reanalysis, and 10% are due to coincident detections.

4.1.2 Introduction

Long records of climate observations are crucial for monitoring regional and global climate change and understanding the underlying climate processes (Trenberth et al., 2013; Dunn et al., 2021). However, many observational climate data are affected by inhomogeneities due to changes in instrumentation, in station location, in observation and processing methods, and/or in the measurement conditions around the station (Peterson et al., 1998a; Mitchell and Jones, 2005; Menne et al., 2009). Inhomogeneities often take the form of abrupt changes, which are detrimental to estimating trends and multi-scale climate variability (Jones et al., 1986; Easterling and Peterson, 1995). Various homogenization methods have been developed for the detection and the correction of such change-points in the context of climate data analysis (Peterson et al., 1998b; Reeves et al., 2007; Costa and Soares, 2009; Venema et al., 2012). The change-point detection step, also called segmentation, can be performed in two classical ways, using either a statistical test (e.g. (Alexandersson, 1986; Easterling and Peterson, 1995; Menne and Williams, 2005; Menne and Williams, 2009; Szentimrey, 2008; Wang et al., 2010)) or a penalized likelihood approach (e.g. (Caussinus and Mestre, 2004; Lu et al., 2010; Domonkos, 2011; Mestre et al., 2013)). The former proceeds sequentially and detects one change-point at a time, which leads inevitably to a sub-optimal solution. On the other hand, the second approach estimates all the change-points at once, and is thus optimal or sub-optimal, depending on the search algorithm. When the whole parameter space is explored, such as with the dynamic programming algorithm, the method is optimal. Many climate segmentation methods are used on differenced data, where the target series is differenced with respect to a reference series. Using differenced series helps to remove the common climate signal and improves thus the detection power of the segmentation method. However, one drawback of this approach is that any detected change-point can be either due to the target series or to the reference series, if the latter is not homogeneous. In this so-called relative homogenization approach, the reference series has been traditionally constructed by compositing the series from several nearby stations (Alexandersson, 1986; Menne and Williams, 2005; Guijarro, 2011). Compositing relaxes the need for homogeneous reference series thanks to the averaging from many nearby stations, such that the detected change-points can be attributed with good confidence to the target series. Unfortunately, in practice, composited reference series often contain non-negligible inhomogeneities. Another approach based on the pairwise comparison of individual series has been shown to be an interesting alternative (Caussinus and Mestre, 2004; Menne and Williams, 2009; Mestre et al., 2013; Domonkos et al., 2021). In this approach, the change-points from the target and reference series are disentangled in a post-segmentation step, referred to as "attribution". In Caussinus and Mestre (2004), the attribution step is done manually, by using both statistical inference and historical information (station metadata) in an iterative way. In Menne and Williams (2009), an automatic procedure is proposed that attributes a change-point to the station with the highest overall count of detections. This method also uses station metadata when available. It assigns the detected change-points to the nearest known event from the station history within some confidence limit. In Mestre et al. (2013), both a semi-automatic method similar to Caussinus and Mestre (2004) and a fully automatic method based on the joint detection of all series at once are implemented, but the latter is not a relative homogenization method and is thus not recommended (Domonkos, 2021).

The above-mentioned attribution methods generally require many nearby stations in order to find out which station is the cause of the detected change-point. They also operate in an iterative way, alternating the segmentation and attribution steps, and perform better when history information is included. In this work, we propose a new attribution method which significantly relaxes these constraints. Firstly, it works even if only one nearby station is available, which makes it usable in data sparse networks. Secondly, it operates in a post-processing mode, meaning that it uses as input the results from the segmentation step and does not need to iterate, although iterations may possibly help to make it more robust. Thirdly, it uses a predictive rule based on machine learning to attribute the cause of change-point among the tested series. The latter is trained in a preliminary step based on real data and is thus optimized for the specific data of interest.

The data of interest in this work is Integrated Water Vapor (IWV) derived from Global Navigation Satellite System (GNSS) measurements (Bock, 2019) and from the fifth ECMWF reanalysis (ERA5) (Hersbach et al., 2020). Because the global GNSS data set is relatively sparse, the reanalysis is used as a reference to form the target minus reference difference series necessary to the segmentation step (Ning et al., 2016; Bock et al., 2019; Van Malderen et al., 2020; Nguyen et al., 2021; Quarello et al., 2022). The primary goal of the attribution is thus to determine whether any change-point detected by the segmentation is due to the GNSS series or to the reanalysis series. Although the long-term stability of reanalyses is sometimes questioned (Thorne and Vose, 2010), they are rather homogeneous, especially in recent years (roughly after 2000) (Sterl, 2004; Kozubek et al., 2020). Inhomogeneities in reanalyses are mainly suspected when changes occur in the global observing system, e.g. the start or end of satellite missions which data are assimilated (Rienecker et al., 2011; Schroeder et al., 2016).

In this study, the GNSS minus reanalysis data are segmented with the "GNSSseg" method developed by Quarello (2020) which is based on a penalized likelihood approach. It detects abrupt changes in the mean in the presence of a periodic (seasonal) bias and a periodic variance (on a monthly basis) and is available on the CRAN (`https://cran.r-project.org/web/packages/GNSSseg/index.html`). It has been used in a benchmark exercise where it was ranked one of the best among 8 segmentation tools (Van Malderen et al., 2020). The attribution step was not necessary in that simulation study because the reference series was homogeneous.

Figure 4.1 helps to explain the idea of the attribution method proposed in this paper. Let us denote by G and E the GNSS and ERA5 reanalysis series of the main station, respectively, and G' and E' those from a nearby station. We denote by t_1, t_2, and t_3, the change-points detected by the segmentation method in the G-E series, and by t'_1 and t'_2, the change-points detected in the G'-E' series. These change-points have jumps in the mean of +1, -1, and -0.5 signal unit for the G-E series, and +1 and -0.5 signal unit for the G'-E' series. Note that in this sketch, the time period of the G'-E' series covers all the change-points of the main station, but in practice, several nearby stations may be necessary. The positions of the change-points illustrate different typical situations encountered in practice with our data. The first change-point in the nearby station, t'_1, is quite far from all the change-points detected in the main station. This illustrates the fact that the causes of inhomogeneities in GNSS data are primarily station-specific, i.e. coincident change-points in G and G' are expected to be rare. On the other hand, t'_2 is close in time to t_3 which illustrates an inhomogeneity in the reanalysis data with a large spatial extension, i.e. impacting both E and E'. Real data often contain data gaps which are due to instrumental failures

leading eventually to an equipment change and possibly to an inhomogeneity. This situation is illustrated with a gap after t_2 in the G-E series. The likeliness of these different situations is summarized in the following "empirical" rules which will help interpreting the features seen in the difference series:

(R1) it is unlikely that change-points in two different GNSS series (here G and G') occur at the same time because they are station-specific in nature (e.g. hardware failure, equipment change, local environmental change).

(R2) on the other hand, it is likely that change-points in the reanalysis occur simultaneously at the main and nearby sites (impacting E and E') because they are expected to have a large spatial extent (e.g. due to a change in assimilation of satellite measurements).

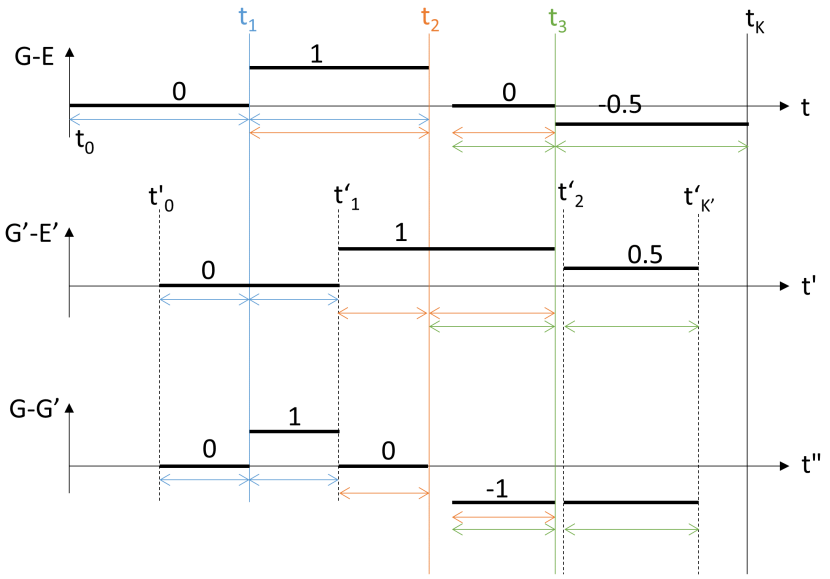


Figure 4.1 – Schematic view of three paired series of differences, G-E, G'-E', and G-G', where G and E are the series from the main station, and G' and E' the series from the nearby station. Change-points detected by the segmentation method in the main (nearby) station are noted t_k (t'_k) and are indicated by the vertical solid (dotted) lines. By convention, t_0 (t'_0) and t_K ($t'_{K'}$) refer to the time of the first and last observation, respectively, in the main (nearby) station. The colored horizontal lines with arrows indicate the segments on the left and the right of the change-points that are used to estimate the deterministic and stochastic parameters of the regression model. This figure is discussed the Introduction.

Inspection of the first two series of differences in Figure 4.1 in the light of these rules suggests that t_1 is likely due to a +1 jump in G, t'_1 is likely due to a +1 jump in G', t_2 is likely due to a -1 jump in G, and t'_2 and t_3 are likely due to a -0.5 jump in both E and in E'. However, to confirm these guesses, we need to inspect additional series of differences combining more of the four base series (G, E, G', and E'). The lower plot in Figure 4.1 shows the G-G' series. It is straightforward, by the same reasoning, to confirm the guessed interpretation of the former two series. In a more general procedure, we would use all six combinations of the four base series and by deduce which of the four base series is/are the cause of the jumps observed in the multiple differenced series.

BS Table

	Jump in the base series				Conditional probability	Joint probability
	G	E	G'	E'	P(G', E' | G, E)	P(G, E, G', E')
1	1	0	0	0	0.8	0.18225
2	1	0	0	1	0.045	0.010125
3	1	0	0	-1	0.045	0.010125
4	1	0	1	0	0.045	0.010125
5	1	0	1	1	0.0025	0.0005625
6	1	0	1	-1	0.0025	0.0005625
7	1	0	-1	0	0.045	0.010125
8	1	0	-1	1	0.0025	0.0005625
9	1	0	-1	-1	0.0025	0.0005625
10	0	-1	0	0	0.045	0.010125
11	0	-1	0	1	0.045	0.010125
12	0	-1	0	-1	0.81	0.18225
13	0	-1	1	0	0.0025	0.0005625
14	0	-1	1	1	0.0025	0.0005625
15	0	-1	1	-1	0.045	0.010125
16	0	-1	-1	0	0.0025	0.0005625
17	0	-1	-1	1	0.0025	0.0005625
18	0	-1	-1	-1	0.045	0.010125
19	-1	0	0	0	0.81	0.18225
20	-1	0	0	1	0.045	0.010125
21	-1	0	0	-1	0.045	0.010125
22	-1	0	1	0	0.045	0.010125
23	-1	0	1	1	0.0025	0.0005625
24	-1	0	1	-1	0.0025	0.0005625
25	-1	0	-1	0	0.045	0.010125
26	-1	0	-1	1	0.0025	0.0005625
27	-1	0	-1	-1	0.0025	0.0005625
28	0	1	0	0	0.045	0.010125
29	0	1	0	1	0.81	0.18225
30	0	1	0	-1	0.045	0.010125
31	0	1	1	0	0.0025	0.0005625
32	0	1	1	1	0.045	0.010125
33	0	1	1	-1	0.0025	0.0005625
34	0	1	-1	0	0.0025	0.0005625
35	0	1	-1	1	0.045	0.010125
36	0	1	-1	-1	0.0025	0.0005625
37	1	-1	0	0	0.045	0.00225
38	1	-1	0	1	0.045	0.00225
39	1	-1	0	-1	0.81	0.0405
40	1	-1	1	0	0.0025	0.000125
41	1	-1	1	1	0.0025	0.000125
42	1	-1	1	-1	0.045	0.00225
43	1	-1	-1	0	0.0025	0.000125
44	1	-1	-1	1	0.0025	0.000125
45	1	-1	-1	-1	0.045	0.00225
46	-1	1	0	0	0.045	0.00225
47	-1	1	0	1	0.81	0.0405
48	-1	1	0	-1	0.045	0.00225
49	-1	1	1	0	0.0025	0.000125
50	-1	1	1	1	0.045	0.00225
51	-1	1	1	-1	0.0025	0.000125
52	-1	1	-1	0	0.0025	0.000125
53	-1	1	-1	1	0.045	0.00225
54	-1	1	-1	-1	0.0025	0.000125

SD Table

	Jump in the series of difference					
	G-E	G-G'	G-E'	E-E'	G'-E'	G'-E
1	1	1	1	0	0	0
2	1	1	0	-1	-1	0
3	1	1	2	1	1	0
	1	0	1	0	1	1
	1	0	0	-1	0	1
4	1	0	2	1	2	1
5	1	2	1	0	-1	-1
6	1	2	0	-1	-2	-1
7	1	2	2	1	0	-1
8	1	0	0	-1	0	1
9	1	0	-1	-2	-1	1
10	1	0	1	0	1	1
11	1	-1	0	-1	1	2
12	1	-1	-1	-2	0	2
13	1	-1	1	0	2	2
	1	1	0	-1	-1	0
14	1	1	-1	-2	-2	0
	1	1	1	0	0	0
15	-1	-1	-1	0	0	0
16	-1	-1	-2	-1	-1	0
17	-1	-1	0	1	1	0
18	-1	-2	-1	0	1	1
19	-1	-2	-2	-1	0	1
20	-1	-2	0	1	2	1
	-1	0	-1	0	-1	-1
21	-1	0	-2	-1	-2	-1
	-1	0	0	1	0	-1
22	-1	0	0	1	0	-1
23	-1	0	-1	0	-1	-1
24	-1	0	1	2	1	-1
	-1	-1	0	1	1	0
	-1	-1	-1	0	0	0
25	-1	-1	1	2	2	0
26	-1	1	0	1	-1	-2
27	-1	1	-1	0	-2	-2
28	-1	1	1	2	0	-2
29	2	1	1	-1	0	1
30	2	1	0	-2	-1	1
31	2	1	2	0	1	1
32	2	0	1	-1	1	2
33	2	0	0	-2	0	2
34	2	0	2	0	2	2
35	2	2	1	-1	-1	0
36	2	2	0	-2	-2	0
37	2	2	2	0	0	0
38	-2	-1	-1	1	0	-1
39	-2	-1	-2	0	-1	-1
40	-2	-1	0	2	1	-1
41	-2	-2	-1	1	1	0
42	-2	-2	-2	0	0	0
43	-2	-2	0	2	2	0
44	-2	0	-1	1	-1	-2
45	-2	0	-2	0	-2	-2
46	-2	0	0	2	0	-2

RSD table

	Restricted jump in the series of difference					
	G-E	G-G'	G-E'	E-E'	G'-E'	G'-E
1	1	1	1	0	0	0
2	1	1	0	-1	-1	0
3	1	1	1	1	1	0
	1	0	1	0	1	1
	1	0	0	-1	0	1
4	1	0	1	1	1	1
5	1	1	1	0	-1	-1
6	1	1	0	-1	-1	-1
7	1	1	1	1	0	-1
8	1	0	0	-1	0	1
9	1	0	-1	-1	-1	1
10	1	0	1	0	1	1
11	1	-1	0	-1	1	1
12	1	-1	-1	-1	0	1
13	1	-1	1	0	1	1
	1	1	0	-1	-1	0
14	1	1	-1	-1	-1	0
	1	1	1	0	0	0
15	-1	-1	-1	0	0	0
16	-1	-1	-1	-1	-1	0
17	-1	-1	0	1	1	0
18	-1	-1	-1	0	1	1
19	-1	-1	-1	-1	0	1
20	-1	-1	0	1	1	1
	-1	0	-1	0	-1	-1
21	-1	0	-1	-1	-1	-1
	-1	0	0	1	0	-1
22	-1	0	0	1	0	-1
23	-1	0	-1	0	-1	-1
24	-1	0	1	1	1	-1
	-1	-1	0	1	1	0
	-1	-1	-1	0	0	0
25	-1	-1	1	1	1	0
26	-1	1	0	1	-1	-1
27	-1	1	-1	0	-1	-1
28	-1	1	1	1	0	-1
29	1	1	1	-1	0	1
30	1	1	0	-1	-1	1
31	1	1	1	0	1	1
32	1	0	1	-1	1	1
	1	0	0	-1	0	1
	1	0	1	0	1	1
33	1	1	1	-1	-1	0
	1	1	0	-1	-1	0
	1	1	1	0	0	0
34	-1	-1	-1	1	0	-1
35	-1	-1	-1	0	-1	-1
36	-1	-1	0	1	1	-1
37	-1	-1	-1	1	1	0
	-1	-1	-1	0	0	0
	-1	-1	0	1	1	0
38	-1	0	-1	1	-1	-1
	-1	0	-1	0	-1	-1
	-1	0	0	1	0	-1

Table 4.1 – BS Table: Theoretical configurations of the jumps in the four base series (G, E, G', E'), coded as: 0=no jump, -1=downward jump, +1=upward jump, and the associated conditional and joint probabilities. SD Table: resulting jumps in the series of differences, coded on five levels (-2, -1, 0, 1, 2). RSD Table: similar to SD Table, but values are restricted to three levels only (-1, 0, +1). In theory the six test results could be searched in the RSD Table and the corresponding configuration of the jumps in the four based series attributed. Duplicated results in the SD and RSD Tables are highlighted with colored background; the configurations of lower joint probabilities are removed, leaving 46 cases in the SD Table and 38 cases in the RSD Table. See Appendix B for further details.

The BS Table in Table 4.1 displays all the relevant combinations of jumps/no jumps in the four base series. The corresponding "theoretical" test results for the series of difference are given in the SD Table. The RSD Table shows the corresponding "practical" test results where the values are restricted to ± 1 (meaning an upward/downward jump) and 0 (meaning no jump, see Appendix A for further details). The latter two tables contain some duplicate configurations highlighted by the colored background which can be distinguished based on the joint probabilities (see Appendix B for the computation of the probabilities). In the end, we can distinguish 38 configurations in the RSD Table. In theory, we could thus use this table to attribute the jumps in the four base series based on the test results of the six series of differences. However, in practice, some of the test results may be wrong due to false negatives and/or false positives, and the combination of the six test results

would not be in the table. To overcome this difficulty, our attribution procedure builds on two main bricks. First, it uses an efficient test based on the Generalized Least Squares (GLS) method. This method is known to have higher detection power than most traditional regression methods in the presence of heteroscedastic and autocorrelated noise as is the case with our data. Second, a predictive rule is constructed based on the tests results from real data, using a machine learning algorithm. This is an efficient way to predict the most likely solution when the combination of the six test results is not in the table.

In subsection 4.1.3, we describe the stochastic properties of our data set composed of IWV time series from ground-based GNSS data and the ERA5 data. We highlight the embedded heteroscedasticity and autocorrelation of these differenced IWV series. In subsection 4.1.4, we evaluate the power of several test methods for testing a fixed change in mean with simulated data mimicking the heteroscedasticity and autocorrelation properties of our real data. We show that the GLS approach is superior to the other methods. In subsection 4.1.5, we describe the method for the construction of the predictive rule, compare the performance of four popular machine learning methods, and present the results on a real data set. Subsection 4.1.6 discussed the results and concludes.

4.1.3 Data characterization

4.1.3.1 Data sets

In this study, we use daily IWV data from GNSS observations and from the ERA5 reanalysis which have been segmented beforehand using the GNSSseg segmentation method (Quarello et al., 2022). Our main goal here is to attribute each of the change-points detected with GNSSseg either to the GNSS series or to the ERA5 series. Several past studies using similar data highlighted the presence of abrupt changes in the mean of GNSS series (Vey et al., 2009; Bock et al., 2014; Ning et al., 2016; Parracho et al., 2018) or in the reanalysis data (Schroeder et al., 2016; Ning et al., 2016; Parracho et al., 2018; Nguyen et al., 2021). Inhomogeneities in the GNSS data are mainly due to equipment changes and changes in the station's environment, but the magnitude of the jumps may also depend on the data processing procedure (Nguyen et al., 2021). In this work, we use reprocessed GNSS data from Center for Orbit Determination in Europe (CODE), covering the period from 1994 to 2014 (REPRO2015), extended until the end of 2018 by a consistent operational processing. Details of the processing are described in Nguyen et al. (2021) and references therein, and the data set is available from Bock (2019). These data are from a global network of 436 stations. Data from the ERA5 reanalysis have been extracted at the location of each station and the difference series, G-E, have been segmented using the GNSSseg package. For the purpose of the present study we selected 81 stations with the longest time series. These will be our "main stations". Nearby stations were searched with a distance limit of 200 km in horizontal and 500 m in vertical, but very few were found in the CODE data set. So we used instead the GNSS data reprocessed by the Nevada Geodetic Laboratory (NGL) which comprises nearly 20 thousand stations (Blewitt et al., 2018). The NGL data were converted to IWV, differenced with respect to ERA5, and passed through the GNSSseg segmentation as well. We ended up with 114 detected change-points in 49 main stations that can be tested with respect to 312 nearby stations, resulting in a total number of 494 main/nearby pairs.

4.1.3.2 Pre-processing

Before we form the six difference series and estimate the change in the mean, the IWV data are adjusted for the station height difference and screened for outliers. The IWV adjustment is done following the method described in Bock et al. (2022) with correction model coefficients estimated from ERA5 on a global grid. This step is important when the main station and nearby station are not at the same altitude. It impacts both the GNSS data and the ERA5 data as the latter are extracted at the height of each GNSS station, either main or nearby. Once the G' and E' data are made consistent with the G and E data, the six difference series are formed.

The screening consists in two steps. The first one is a classical outlier detection procedure in which the data points exceeding three standard deviations from the median are removed. In this procedure, the median and a robust standard deviation estimator (Yohai and Zamar, 1988) are computed in a sliding window of length +/- 60 days around the current point. To guarantee the representativeness and accuracy of the estimates in the presence of data gaps, a minimum number of 20 data points are required.

The second step consists in removing data in short segments (less than 80 days) pertaining to a cluster of change-points detected in the main station (see an example in Figure 3 in Quarello (2020)). This problem occurs occasionally in regions where the GNSS data and reanalysis data have a significant representativeness difference (Bock and Parracho, 2019). In such situation, we keep the first change-point but remove data between the the first and the last change-points of the cluster. Similarly, when a change-point in the nearby station is very close (less than 10 days) to a change-point in the main station, we consider that they are both due to the same cause (most likely a change in the reanalysis) and we remove the data points in the nearby series between the two change-points. This case is illustrated in Figure 4.1 where the data between t_3 and t_2' have been removed. In Figure 4.1, we also illustrate the case of a gap in the G-E series just after t_2 which may be due to a screened cluster. Note that, as a result of the screening, the number of data points in a difference series combining the main and nearby sites (e.g. G-G') is always less or equal to that of a collocated difference series (e.g. G-E or G'-E'). In the proposed test procedure, a minimum number of 200 consecutive points is required on each side of the change-points.

4.1.3.3 Characterization of the data and model building

The GNSS minus reanalysis difference series show usually strong heteroscedasticity and periodic (seasonal) biases, along with weak autocorrelation (Quarello et al., 2022). In the following, a series is modelled using the following regression model:

$$z_t = \mu_L + \delta x_t + s_t + e_t, \tag{4.1}$$

where t refers to the time, μ_L is the mean of the signal on the left of the change-point, δ is the amplitude of the jump, x_t is a step function ($x_t = 0$ if $t \leq t_k$ and 1 if $t > t_k$, where t_k is the time of the change-point detected by the segmentation method), s_t is the Fourier series, and e_t is the noise term. For ease of notation, we use t as the time index, with $t = 1, ..., n$, but in reality the data may contain gaps and the time values are not consecutive. To account for this, t can be replaced by $t(i)$, with $i = 1, ..., n$. To account for both heteroscedasticity and autocorrelation, we follow José C. Pinheiro (2000) and represent e_t as the product of two factors:

$$e_t = e_t^* \sigma_t, \tag{4.2}$$

where e_t^* represents a stationary autocorrelated process of unit variance and σ_t^2 is the time-varying variance of e_t, i.e. $var[e_t] = \sigma_t^2$. Preliminary investigation of our data showed that most of the time the noise model is well approximated by an AR(1). Other noise models such as MA(1), ARMA(1,1), and pure white noise occur sometimes. We tested also for higher order ARMA(p,q) models and they are very rare. We limit thus ourselves to the four possible ARMA(p,q) models, with $p, q \in \{0, 1\}$. Recall that an ARMA(1,1) model writes (Shumway and Stoffer, 2017):

$$e_t^* = \phi e_{t-1}^* + \theta w_{t-1} + w_t, \tag{4.3}$$

where w_t is a Gaussian white noise. The noise model identification and parameter estimation methods are described in the next subsection. Note that other stochastic models including periodic variations in the mean, heteroscedasticity and autocorrelation have been proposed by Lund et al. (1995).

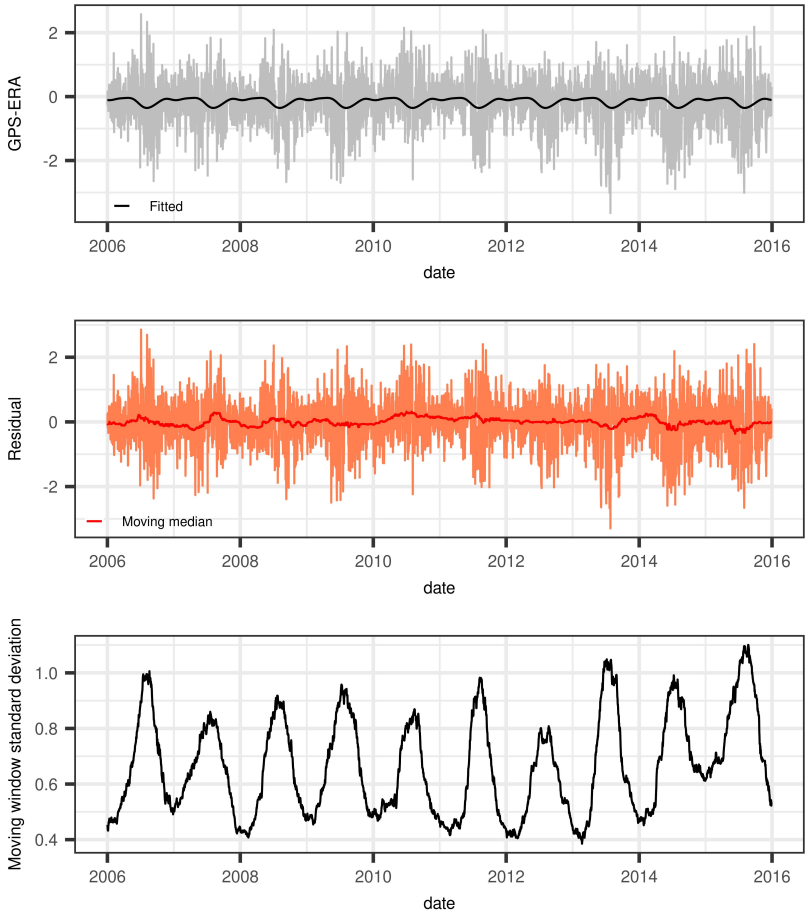


Figure 4.2 – Top: GNSS minus ERA5 time series at station ALBH (Victoria, Canada), in gray, and estimated Fourier series, in black, for a long, homogeneous, segment (no change-point detected by the segmentation method). Middle: FGLS regression residuals (jagged curve) and moving median (smooth curve). Bottom: moving standard deviation illustrating the strong heteroscedasticity in the data.

Figure 4.2 shows an example of a time series (jagged gray curve), with the estimated Fourier series (smooth black curve), the estimated standard deviation (SD), $\hat{\sigma}_t$ (black curve at bottom), and the regression residuals (jagged orange curve). The strong heteroscedasticity is obvious, and because it is not stationary, we used a moving window approach (similar to the outlier screening procedure described above) to estimate it.

Table 4.2 and 4.3 summarize the characteristics of our data set in terms of heteroscedasticity and noise structure, respectively, for all six series of difference (G-E, G-G'...), for the main and all nearby stations. The results are sorted according to the distance between the main and the nearby stations (smaller or larger than 50 km). Regarding the heteroscedasticity, three groups can be identified when the distance between sites is small. The first group (G1) includes G-E and G'-E', i.e. the series with collocated data, which have moderate mean SD of 0.7 kg m^{-2}. The second group (G2) includes E-E' and G-G', i.e. the series comparing the same technique, which have the smallest mean SD (0.5 kg m^{-2}). The last group (G3) involves data from non-collocated data and mixed techniques, and gets the largest mean SD. As the distance increases, the mean SD of series involving different sites increases, as expected from increased representativeness differences. Another striking feature is that the half-range of the variation in SD is around 70% for all six series, indicating that heteroscedasticity is a strong feature in our data.

	Mean of SD		Half-range of SD (%)
Distance	< 50 km	≥ 50 km	
G–E	0.7 ± 0.26		72 ± 20
G'–E'	0.66 ± 0.24		67 ± 19
G–G'	0.52 ± 0.17	1.31 ± 0.47	63 ± 21
E–E'	0.41 ± 0.17	1.26 ± 0.47	73 ± 26
G–E'	0.82 ±0.21	1.38 ± 0.46	67 ± 21
G'–E	0.83 ± 0.26	1.39 ± 0.46	66 ± 20

Table 4.2 – Characterization of the heteroskedasticity in the real data from 494 main/nearby series. The table reports the mean and the half-range of the standard deviation for each of the six paired difference series. The mean values are sorted by distance.

Distance	< 50km				≥ 50km			
series	AR(1)	MA(1)	ARMA(1,1)		AR(1)	MA(1)	ARMA(1,1)	
Coefficients	phi	theta	phi	theta	phi	theta	phi	theta
G-E	0.30	0.00	0.57	-0.32				
G'-E'	0.31	0.22	0.59	-0.34				
G-G'	0.33	0.19	0.65	-0.31	0.30	0.22	0.11	0.12
E-E'	0.31	0.21	0.34	0.23	0.29	0.20	0.25	0.20
G-E'	0.33	0.24	0.59	-0.24	0.31	0.21	0.29	0.08
G'-E	0.32	0.21	0.57	-0.28	0.30	0.22	0.18	0.21

Table 4.3 – Characterization of the autoregressive noise structure of the real data. The table reports the mean estimated coefficients of the noise model for each of the six paired difference series, sorted by distance.

Figure 4.3 shows the distributions of the noise models and of the estimated coefficients for the six differences, again sorted according to the distance. The AR(1) model is the dominant model, with a proportion between 50% and 80%, independently of the distance, while the white noise model is extremely rare. The proportion of MA(1) and ARMA (1,1) depends on the distance and the series: ARMA(1,1) is dominant for the collocated series (similar to the noise group G1), as well as for the series comparing the same technique (group G2), when the distance is small. On the opposite, when the distance is large, MA(1) becomes more frequent, like for the series mixing techniques and sites (group G3). The increase of the distance does thus not only increase the variance of the noise but changes also its nature. Another interesting aspect is the values of the coefficients. For the AR(1), they are very similar (around 0.3) for all series, regardless of the distance. Similarly, for the MA(1), they are very similar (around 0.2). More surprisingly, the estimated coefficients of the ARMA(1,1) model for the non-collocated series depend somewhat on the distance, with the exception of E-E'. Values of $\hat{\phi}$ and $\hat{\theta}$ are around 0.6 and -0.3, respectively, for the collocated series and when the distance is small, and around 0.2 for both coefficients, when the distance is large. For E-E', the values are always around 0.2. The ARMA(1,1) models with coefficients of opposite sign found at short distance suggest that in these cases the noise is a mixture of AR(1) and white noise (Shumway and Stoffer, 2017). When the distance increases, the moving average part becomes more important, which may be interpreted as a spatial/temporal averaging of the variability in the difference series. The mean values of the estimated coefficients are reported in Table 4.3.

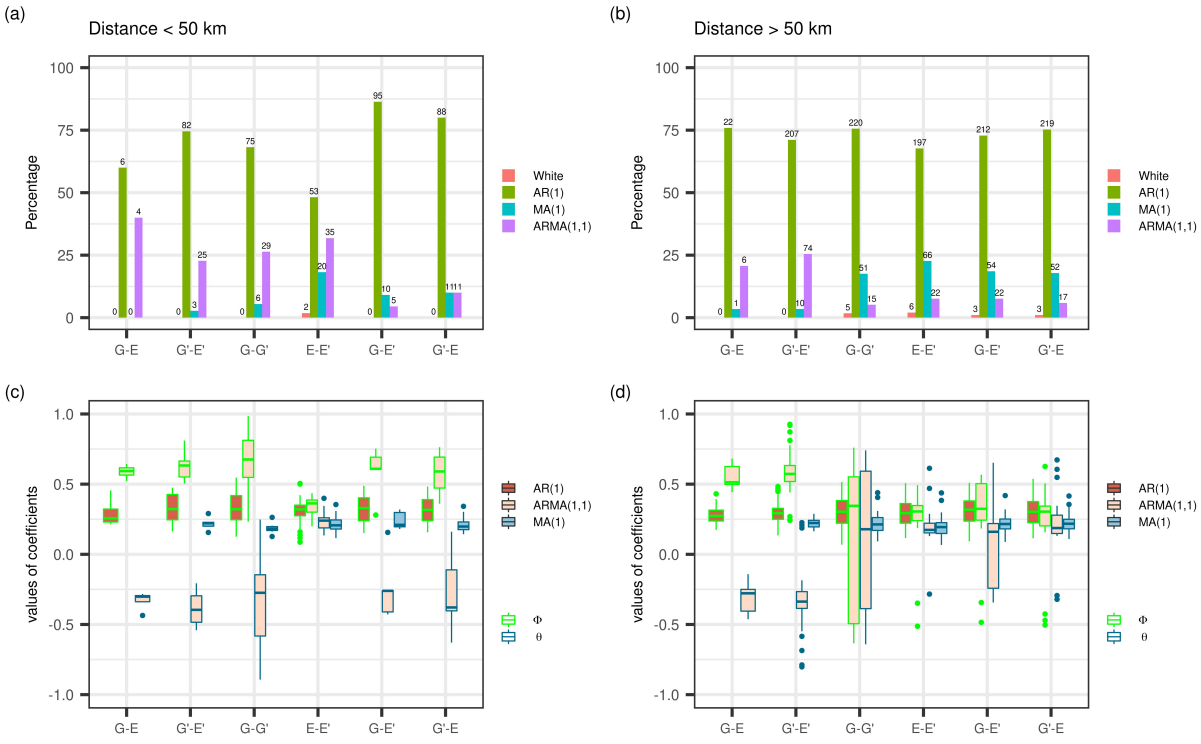


Figure 4.3 – Results of noise model identification in the real data. (a, b) Histogram of model types (white noise, AR(1), MA(1), ARMA(1,1)) selected with `auto.arima` function for each of the six series of differences (G-E, G-E', etc.); the bar heights show the percentage (y-axis) of cases for each series (x-axis), the number of cases is indicated on the top of each bar. (c, d) noise model coefficients, $\hat{\phi}$ and $\hat{\theta}$, estimated with `arima` function, for each model. Results are sorted according to the distance between the main and the nearby stations, (a, c) smaller than 50 km, (b, d) larger than 50 km.

4.1.4 Proposed tests for a fixed change-point

4.1.4.1 Regression model and different tests

In this subsection, we compare different procedures to test the significance of the jumps associated to change-points detected by the segmentation (i.e. considered here at known position). The series modelled by Eq. (4.1), with specifications (4.2) and (4.3), is now rewritten in matrix form:

$$\mathbf{z} = X\beta + \boldsymbol{e}, \tag{4.4}$$

where β includes the coefficients of the deterministic part of the model, $\beta = (\mu_L, \delta, a_1, ..., a_4, b_1, ..., b_4)'$, and X includes the corresponding regressors. Here, the a_l and b_l, $l = 1, ..., 4$, are the coefficients of a Fourier series of order 4, and the corresponding regressors are $\cos(2\pi lt(i)/T)$ and $\sin(2\pi lt(i)/T)$, with $T = 365$ days, and $t(i)$ is the time of the i^{th} observation, z_i, $i = 1, ..., n$. The noise vector, $\boldsymbol{e}$ is assumed to be distributed as $\mathcal{N}(0, \Sigma_0)$, where Σ_0 is the variance-covariance matrix describing the noise model. A correct specification of this matrix is crucial for the subsequent tests of the estimated coefficients of the deterministic model.

When Σ_0 is known, either Ordinary Least Squares (OLS) or Generalized Least Squares (GLS) methods can be used, for which the solutions for $\hat{\beta}$ and the corresponding variance-covariance matrices write:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'\mathbf{z} \quad \text{and} \quad var[\hat{\beta}_{OLS}] = (X'X)^{-1}(X'\Sigma_0 X)(X'X)^{-1}, \tag{4.5}$$

$$\hat{\beta}_{GLS} = (X'\Sigma_0^{-1}X)^{-1}X'\Sigma_0^{-1}\mathbf{z} \quad \text{and} \quad var[\hat{\beta}_{GLS}] = (X'\Sigma_0^{-1}X)^{-1}. \tag{4.6}$$

Recall that, in the presence of heteroscedasticity and/or autocorrelation, GLS is the Best Linear Unbiased Estimator (BLUE) while OLS solution is not, although OLS remains unbiased.

In practice, Σ_0 is typically unknown and needs to be estimated. In the classical linear model (CLM) framework it is generally assumed that the data are independent and homoscedastic, e.g. $\Sigma_0 = \sigma_0^2 I_n$, where I_n is the identity matrix. An estimator of $var[\hat{\beta}_{OLS}]$ is then simply

$$\widehat{var}[\hat{\beta}_{OLS}]_{CLM} = \hat{\sigma}_0^2(X'X)^{-1}. \tag{4.7}$$

where $\hat{\sigma}_0^2 = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/(n-k)$ is an unbiased estimator of the noise variance σ_0^2.

The CLM assumptions are not satisfied with our data and, despite the OLS solution (4.5) remains unbiased, the variance estimator is strongly biased and leads to significant inference errors. To solve this problem, some methods have been proposed in the literature. The two main ones, which we considered in this work, are:

— the so-called OLS-HAC that consists in still using the OLS solution but to consider a consistent estimator for the variance that is the Heteroscedasticity and Autocorrelation Consistent estimator (HAC) (White, 1980; Newey and West, 1986). This estimator is robust to the presence of heteroscedasticy and serial correlations of unknown form and has good asymptotic properties. The key idea is to estimate $M = X'\Sigma_0 X$ instead of Σ_0, which is difficult to estimate due to its large size (it contains nominally $n(n+1)/2$ parameters).

The variance estimator writes:

$$\widehat{var}[\hat{\beta}_{OLS}]_{HAC} = (X'X)^{-1}\hat{M}(X'X)^{-1}. \tag{4.8}$$

A large class of HAC estimators is the non-parametric kernel estimators. A drawback of this type of estimator is that its performance varies with the choices of the kernel function and its bandwidth, but the advantage is that it does not require to specify the covariance structure and is computationally very fast. Here we consider the "Quadratic Spectral" kernel (Andrews, 1991) with its proposed optimal bandwidth value. This method is available the R package `sandwich` (Zeileis, 2006).

— the Feasible GLS (FGLS) which consists in estimating Σ_0, assuming it has a specific structure with a reduced number of parameters to be estimated, in the GLS solution Eq. (6). Denoting $\hat{\Sigma}_n$ the estimator of Σ_0, Eq. (6) becomes:

$$\hat{\beta}_{FGLS} = (X'\hat{\Sigma}_n^{-1}X)^{-1}X'\hat{\Sigma}_n^{-1}\mathbf{z} \quad \text{and} \quad \widehat{var}[\hat{\beta}_{FGLS}] = (X'\hat{\Sigma}_n^{-1}X)^{-1}. \tag{4.9}$$

Following José C. Pinheiro (2000), we decompose Σ_0 as $\Sigma_0 = \boldsymbol{V}\boldsymbol{C}\boldsymbol{V}$, where $\boldsymbol{V} = diag(\sigma_t)$ and $\boldsymbol{C}$ is the correlation matrix associated to the noise model (see Shumway and Stoffer (2017) for the formulation of matrix $\boldsymbol{C}$ as a function of ϕ and θ for an ARMA(1,1)). With this parameterization, Σ_0 is described by a maximum of $n + 2$ coefficients.

Since both $\hat{\beta}_{FGLS}$ and $\widehat{var}[\hat{\beta}_{FGLS}]$ depend on $\hat{\Sigma}_n$, an iterative procedure is implemented. In order to stabilize the convergence and also to avoid over-fitting, we first select the best noise model among the four possible models (ARMA(p,q) with $p, q \in \{0, 1\}$) using the `auto.arima` function in the R package `forecast` (Hyndman and Khandakar, 2008). The BIC criterion is used and is complemented by a test of the significance of the final model coefficients. Next, we use the following iterative procedure:

1. fit the OLS solution (4.5);
2. compute the noise variance $\hat{\boldsymbol{V}}_n = diag(\hat{\sigma}_t^2)$ from the OLS residuals in a moving window (see subsection 4.1.3.3);
3. fit a preliminary FGLS solution (4.6) where Σ_0 is replaced by $\hat{\boldsymbol{V}}_n$;
4. fit the ARMA model coefficients, $\hat{\phi}$ and $\hat{\theta}$, from the FGLS residuals;
5. compute $\hat{\Sigma}_n = \hat{\boldsymbol{V}}_n\hat{\boldsymbol{C}}_n\hat{\boldsymbol{V}}_n$, where $\hat{\boldsymbol{C}}_n$ is the correlation matrix of the ARMA model with the fitted coefficients;
6. fit the final FGLS solution (4.9);
7. repeat steps 3 to 6 until convergence.

In step 4, the parameters $\hat{\phi}$ and $\hat{\theta}$ of the ARMA noise structure are estimated by Maximum Likelihood with the function `arima` in R. In step 7, the convergence is tested from the difference of $\hat{\beta}$, $\hat{\phi}$, and $\hat{\theta}$, of two successive iterations. The maximum number of iterations is set to 10 and is never reached. For

the preliminary selection of the noise model, a simplified scheme including steps 1 to 4 is used, and `auto.arima` is used in step 4 instead of `arima`. Numerical simulations showed that `auto.arima` is able to select the correct model in 99 % of the cases when it is white noise, 93 % when it is MA(1), 90% when it is AR(1), and 63 % when it is ARMA(1,1), for a sample size $n \geq 1000$ with typical values for the coefficients (Figure 4.3).

Note that, compared to the OLS-HAC approach, the FGLS procedure, due to its iterative scheme, is computational time-demanding. However, when the noise model is correctly specified, the FGLS estimate of $\widehat{var}[\hat{\beta}]$ is more accurate than its OLS-HAC counterpart, and the power of the test is improved.

In the following we are interested in testing the null hypothesis $H_0 : \delta = 0$. The associated test statistic is:

$$\tau_{\delta,\text{OLS-HAC}} = \frac{\hat{\delta}_{\text{OLS}}}{\hat{\sigma}_{\hat{\delta}_{\text{HAC}}}} \quad \text{for the OLS-HAC, and} \quad \tau_{\delta,\text{FGLS}} = \frac{\hat{\delta}_{\text{FGLS}}}{\hat{\sigma}_{\hat{\delta}_{\text{FGLS}}}} \quad \text{for the FGLS,} \tag{4.10}$$

where $\hat{\sigma}_{\hat{\delta}_\bullet}$ is the estimated standard error of $\hat{\delta}$ which is extracted from $\widehat{var}[\hat{\beta}_\bullet]$, and $\bullet$ = OLS-HAC or FGLS. For the OLS-CLM and GLS estimators considered in the simulation study, similar statistics are computed using their respective estimators for $\hat{\delta}$ and $\hat{\sigma}_{\hat{\delta}}$ given by Eqs. (5) and (6).

In contrast to the HAC estimator, the asymptotic properties (unbiasedness and efficiency) of the FGLS estimator are typically not known. Numerical simulations with the FGLS procedure described above show that $\hat{\beta}_{FGLS}$ is not biased and its variance is very close to GLS, although slightly larger, but still smaller than HAC. Finally, the distributions of both statistics are very close to $\mathcal{N}(0,1)$ under H_0. The critical value, $\tau_{\alpha/2}$, associated to a given significance level, α, will thus be computed using the standard normal distribution.

4.1.4.2 Evaluation based on simulations

We conducted a large number of simulations, for different types of noise characteristics and sample sizes, to assess the test methods introduced above. In these simulations, we modelled the noise heteroscedasticity by a "raised cosine" function, $\sigma_t^2 = \sigma_m^2 - \sigma_v^2 \cos 2\pi t/T$, where σ_m^2 and $\sigma_v^2 \leq \sigma_m^2$ represent the mean and half-range modulation of the variance over one period, T, respectively.

The results are compared in Figure 4.4 for significance level, $\alpha = 0.05$. It is seen that the False Positive Rate (FPR) stays generally fairly close to the nominal significance level for GLS, FGLS, and OLS-HAC, for most autocorrelation and heteroscedasticity characteristics, except when the autocorrelation is very strong. In contrast, the OLS-CLM method performs very badly when the data is autocorrelated or heteroscedastic, due to the bias in its variance estimator (Eq. 4.7). The power of the test is measured by the True Positive Rate (TPR). It depends on the jump amplitude, the sample size, and the noise characteristics, in addition to the fixed significance level. When ϕ increases, it is generally observed that the TPR decreases, for all four methods. This is due to an increase in $\hat{\sigma}_{\hat{\delta}}$ which is sometimes interpreted as a reduction of the "equivalent sample size" (Zwiers and Storch, 1995). The higher TPR of OLS-CLM is actually a consequence of its higher FPR and does not indicate a good performance per se. In contrast, when the heteroscedasticity increases at constant ϕ, GLS and FGLS clearly outperform OLS-HAC. For OLS-HAC the TPR remains actually constant while for GLS and, to a lesser extent FGLS, the TPR increases with stronger heteroscedasticity. This is explained by the fact that GLS

and FLGS take the weight of every observation into account; when heteroscedasticity is strong, observations with small errors have high weight which leads to a decrease in $\hat{\sigma}_{\hat{\delta}}$. When both heteroscedasticity and autocorrelation are strong, the power of the test is low for all methods, but FLGS still performs better than OLS-HAC.

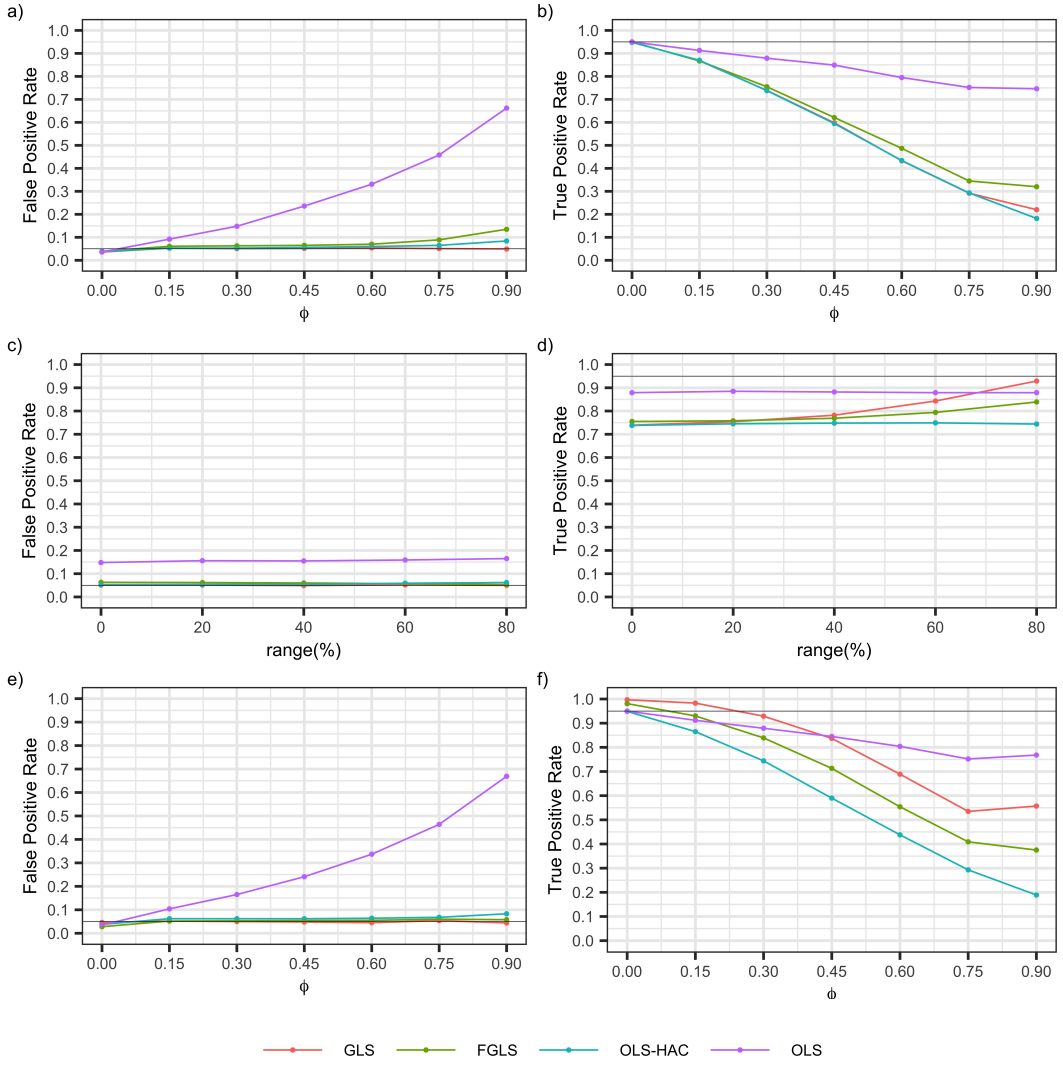


Figure 4.4 – False Positive Rate (FPR) and True Positive Rate (TPR) of jump detections with the regression model (1) and four different estimation methods, for three scenarios: (a, b) AR(1) noise of unit variance with $\phi = 0, ..., 0.90$; (c, d) heteroskedastic and AR(1) noise, with $\phi = 0.3$ and half-range of variance from 0 to 80 %; (e, f) heteroskedastic and AR(1) noise, with $\phi = 0, ..., 0.90$, and half-range of variance of 80 %. For TPR, the amplitude of the jump is fixed to 0.356, which corresponds to TPR=0.95 when $\phi = 0$. The sample size is n=400 and the number of simulated series is m=1000.

Figure 4.5 provides additional insight into the power of the FGLS test as a function of the the jump amplitude and sample size, for a typical noise configuration. A 75 % probability of detection is expected for jumps of 0.25 or larger with a sample size of n=600. Most of the G-E, G-E', and G-G' series in the real data actually fulfill this requirement (see Figure 4.6, further discussed in the next subsection). For stronger noise (usually due to larger distance) or smaller jumps, a larger sample size would be required to maintain a high detection probability.

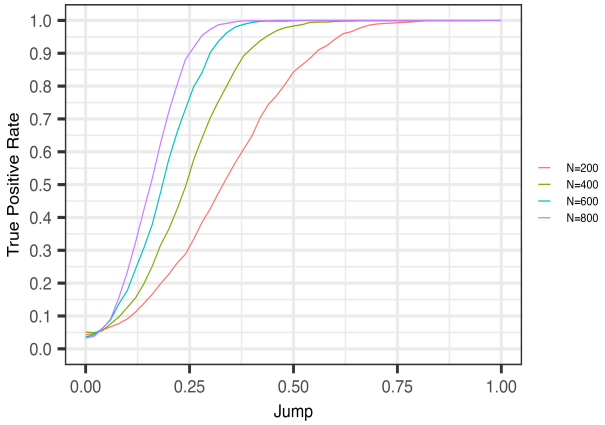


Figure 4.5 – True Positive Rate (TPR) of jump detections with the FGLS method, as a function of jump amplitude and sample size, in case of heteroskedastic and AR(1) noise with $\phi = 0.3$, mean variance of 1, and half-range of variance of 80%.

4.1.4.3 Application to real data

The FGLS procedure was applied to the series of differences from the 494 main/nearby pairs.

Figure 4.6 shows the distribution of estimated jump amplitudes, their standard errors, and the associated absolute t-values computed from Eq. 4.10, where the non-collocated series (G-G', G-E', E-E', and G'-E) are sorted by distance. Notably, the three series involving G have significantly larger median jump amplitudes (around 0.3 kg m^{-2}) than the other three series, regardless of the distance. This result suggests that large jumps are occurring more often in the G series than in the E, E', or G' series. In G'-E' and G'-E, the median jump is small, as expected and expressed in our first rule stating that it is unlikely to have a coincident change-point in a nearby GNSS station when there is one detected in the main station. Additionally, a notable observation is that the median jump in E-E' is much larger at larger distance, which may be due to errors in the estimated jumps induced by a increased noise at larger distance. The variation of the SD of the noise with distance directly impacts also the jump standard error. The standard error of estimated jumps is notably smaller in collocated series, such as G-E and G'-E', as well as in non-collocated series at short distance. Furthermore, the standard errors in G-E' and G'-E (non-collocated series from different techniques) are slightly larger than in G-G' and E-E' (non-collocated series from the same technique) even at short distance, as also noticed in the three noise groups discussed in subsection 4.1.3.3. Finally, the t-values can be interpreted by considering the jump magnitudes and their standard errors. It is evident that the three series involving G yield larger t-values due to higher jump magnitudes. In contrast, the other three series have much smaller t-values, mainly because some of the large jumps at larger distance are damped by the larger standard errors. A common feature to all non-collocated series is that the t-values decrease with distance.

Figure 4.7 shows the corresponding test results with a significance level α=0.05. At short distance, the three series involving G are almost all significant. Especially, all the G-E jumps are significant, which demonstrates a high consistency between our FGLS tests and the segmentation results. Almost all G-E' jumps are significant as well, while almost all E-E' are not significant. This latter result confirms our second rule (E and E' are expected to be consistent, i.e. either no jump or a jump in both, simultaneously). Most G-G' jumps are also

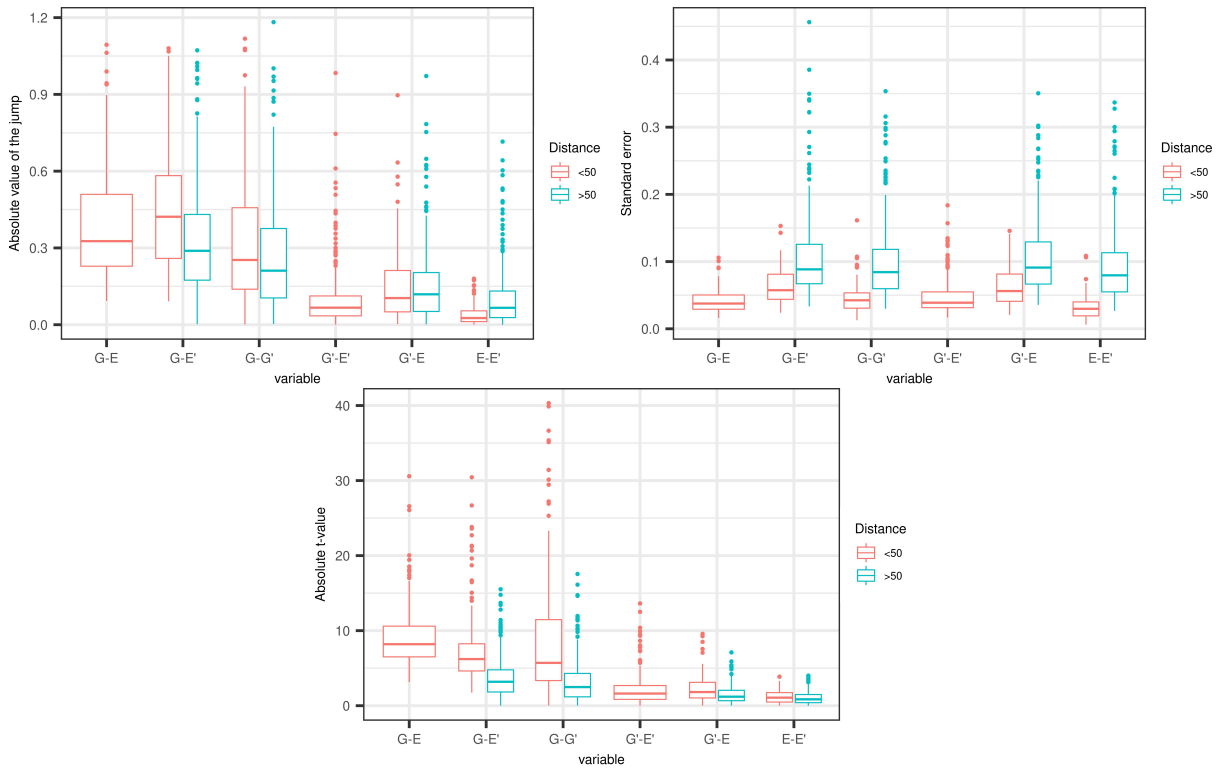


Figure 4.6 – Distribution of absolute jump amplitudes and their standard errors, and the associated t-values computed from the FGLS estimates of the real data (494 main/nearby pairs). The results are sorted based on the main/nearby distance (< 50 km and ≥ 50 km).

significant, which confirms the idea that most jumps are in G. Finally, the G'-E' and G'-E jumps are most of the time insignificant, which again supports of our first rule (G and G' are unlikely to change simultaneously). As the distance increases, the proportion of insignificant jumps also increases due to higher standard errors.

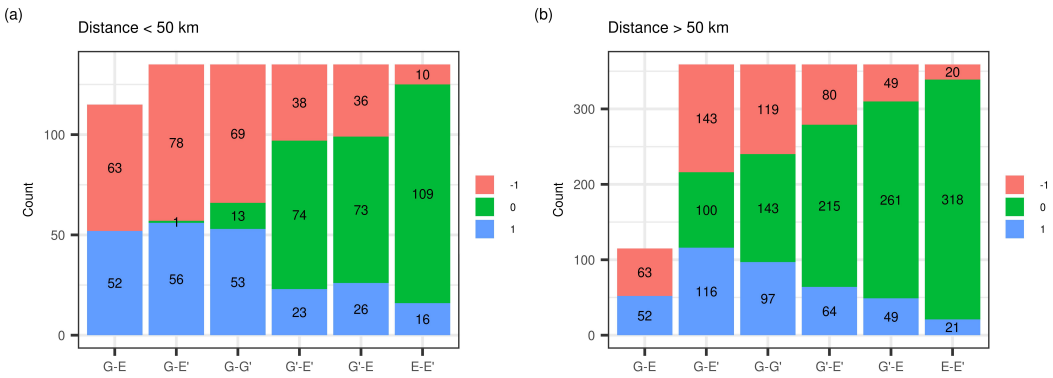


Figure 4.7 – Distribution of test results associated to the estimated amplitudes of jumps shown in Figure 4.6, sorted by distance: (a) < 50 km, (b) ≥ 50 km. Test results are color coded as: green for insignificant and red/blue for significant downward/upward jump.

4.1.5 Predictive rule

The objective is to build a classifier $\widehat{\psi}(x)$ representing the prediction of the configuration y, i.e. the quadruplet composed of G, E, G', and E', given x, the vector composed of the test statistics from the series of differences. In

the development of this classifier, we are confronted with two principal challenges. First, the true configurations are unknown, resulting in the unavailability of y for training, evaluation and prediction. Second, the presence of all configurations in the data is nearly improbable. This is due to the rarity of occurrence for certain configurations according to the probabilities indicated in Table 4.1. To address these challenges, we propose to generate a synthetic dataset based on the $N = 494$ test results of the real data using a bootstrapping technique as introduced by (Efron and Tibshirani, 1993). This dataset would ensure each configuration is represented through a set of (x, y) pairs. We then compare the performance of four popular classifiers for the prediction objective. The reader can refer to subsection 2.2.4 of Chapter 2 for more details about the supervised classification, two resampling methods and a presentation of the four algorithms we considered here.

4.1.5.1 Preliminary considerations

Considered test results. In this task, we employed the test statistics from five series: G-G', G-E', E-E', G'-E', and G'-E', given that the considered cases pertain exclusively to situations where the G-E test holds significance. Denote by $z_\ell = (z_{\ell 1}, \ldots, z_{\ell 5})$ the vector of the five test statistics for the ℓth test in the five series of difference and by $Z = (z_\ell)_{\ell=1,\ldots,N}$ the formed data set of size $N \times 5$, with $N = 494$, which will be called the original data set in the sequel. Note that the results of all the nearby stations for a given change-point in a given main station can be viewed as replicates reducing the real information to 114, i.e. the number of couples of (main station, change-point); the available information is thus quite small.

Considered configurations. In Table 4.1, among the 38 configurations of the RSD Table, there are two doubles of the five coded test results with same prior probabilities: configurations (7,28) and (12,19). We decide to keep the configurations 7 and 19 which contains a change-point in G. This reduces the total number of configurations to $C = 36$.

The four considered learning algorithms are: the linear discriminant analysis (LDA) (Fisher, 1936), the classification and regression trees (CART) (Breiman et al., 1984), the Random Forest (RF) (Ho, 1995) and the k Nearest Neighbors (k-NN) (Fix and Hodges, 1989; Cover and Hart, 1967). The latter three involve parameters that need to be tuned. They are here automatically optimized by K-fold cross-validation with $K = 10$ using the generic function 'train' of the R package `caret`.

Building of the complete synthetic data set. As mentioned earlier, the bootstrapping technique has been employed to construct the synthetic dataset, operating on the principle of random sampling with replacement from the original data. More precisely, for each configuration y and each series of difference j, we create N_y vectors of the five test statistics or t-values (the sample x) by resampling among the test statistic values $(z_{\ell j})_\ell$ that lead to the test conclusion of y. The correspondence is made with respect to the test outcome ($-1, 0$ or 1) for a given significance level α. As an example, let's consider the configuration 1 in Table 4.1. Each test statistic (t-value) is randomly selected from the respective series of difference, ensuring that the significance levels of these five t-values are $(1, 1, 0, 0, 0)$. The constructed data set is noted $D = \{(y_\ell, x_\ell)_\ell\}_{\ell=1,\ldots,n}$ of size $n = \sum_y N_y$.

A notable consideration is the potential severe imbalance in the configurations within the data, as discussed in subsection 2.2.4.3 of Chapter 2. This feature is well-known to produce biased classifiers for the minor

configurations. Two options can be considered: if we consider the imbalanced aspect as a problem, we can use the same number of replicates $N_y = R$ for each configuration y. This case is called the 'balanced sample case'. On the contrary, if we want to take into account this reality, we can take a number of replicates proportional to the prior probability of each configuration given in Table 4.1, i.e. $N_y = np_y$ if n is the total number of data. This case is called the 'imbalanced sample case'.

4.1.5.2 The proposed Cross-Validation Bootstrap (CVB) procedure

Cross-validation is a popular statistical technique to test a classifier. It involves splitting the data into two subsets: the learning set, on which the classifier is constructed, and a test set, on which the classifier is tested. Since observations of the complete data set D are replicated from the original data set Z, which is small and repeated, there is a risk of overlap between the learning and test data set, inducing inevitably a bias and leading to an underestimation of misclassification error. This is why, we propose here a so-called cross-validation bootstrap (CVB) strategy which consists in first splitting the original data set Z into the learning and test subsets before constructing the complete data set D. The proposed CVB procedure is described in Algorithm 1.

Data: the original data Z

for $b = 1$ *to* B **do**

1. sample a learning data set $Z^{b,L}$ from Z with probability 0.8, and form the test data set $Z^{b,T}$ with the remaining 20% of data. The random sampling is performed on the rows of Z, i.e. on each test;
2. form the two associated complete data sets $D^{b,L}$ and $D^{b,T}$ from $Z^{b,L}$ and $Z^{b,T}$ by preserving the learn/test proportion of 80%/20%, i.e. for each configuration y, $D^{b,L}$ contains $0.8N_y$ samples, and $D^{b,T}$ $0.2N_y$. In the 'balanced sample case', we chose $N_y = R = 100$ and in the 'imbalanced sample case', the smallest value N_y is chosen to 5, leading to a learn sample containing 4 data and a test sample containing only one data;
3. construct the four classifiers on the learning data set $D^{b,L}$: $\widehat{\psi}^{b,k}$, $k \in$ LDA,CART,RF,k-NN;
4. compute the misclassification error of the classifiers on the test data set $D^{b,T}$ with n_T rows:

$$\mathrm{err}^{b,k} = \sum_{\ell=1}^{n_T} \mathbb{1}_{\{\widehat{\psi}^{k,c}(x_\ell^{b,T}) \neq y_\ell^{b,T}\}} \qquad \text{for } k \in \text{LDA, CART, RF, k-NN}$$

end

Evaluation: compute the mean and the standard deviation of misclassification error for each classifier, e.g. the mean is given by:

$$\overline{\mathrm{err}}^k = \frac{1}{B}\sum_{b=1}^{B} \mathrm{err}^{b,k} \qquad \text{for } k \in \text{LDA, CART, RF, k-NN}$$

Algorithm 1: The CVB procedure.

Table 4.4 gives the mean and the standard deviation of misclassification error for the four considered classifiers with $B = 20$. The table presents results for three scenarios: the first and second one involve constructing the

complete dataset using different sampling (balanced vs. imbalanced), both with $\alpha = 5\%$, while the last one employs a balanced sampling with $\alpha = 1\%$. Compared to the balanced sample, the misclassification error is lower for the imbalanced sample in the case of LDA and k-NN, but slightly higher for CART and RF. Similar behavior is observed when comparing learning with $\alpha = 5\%$ and $\alpha = 1\%$ for the balanced sample. Overall, the Random Forest algorithm outperforms the other classifiers in all three scenarios, with the best performance achieved when trained with a balanced sample with $\alpha = 5\%$. We thus choose the Random Forest algorithm and select as the final predictive rule, $\widehat{\psi}$, the one with smallest error among the B classifiers. The predictive power of the five series of difference based on the accuracy criterion (the percentage of correct predictions) are in the decreasing order: E-E', G-G', G-E', G'-E and G'-E'.

c	test level	sample case	LDA	CART	KNN	RF
$\overline{\text{err}}^c$	0.05	balanced	0.1463 ± 0.021	0.0142 ± 0.011	0.1412 ± 0.018	0.0049 ± 0.003
	0.05	imbalanced	0.1108 ± 0.004	0.0165 ± 0.010	0.0351 ± 0.004	0.0054 ± 0.004
	0.01	balanced	0.1424 ± 0.029	0.0210 ± 0.0417	0.1301 ± 0.022	0.0106 ± 0.033

Table 4.4 – Mean misclassification error ± one standard deviation, for the four classifiers in three scenarios: 'balanced sample' with $N_y = R = 100$ and $\alpha = 0.05$, 'balanced sample' and $\alpha = 0.01$, 'imbalanced sample' with $N_y = np_y$ and $\alpha = 0.05$.

4.1.5.3 Application to the real data set

The objective is to predict the configuration for each change-point of each main station. When several nearby stations are available for a given change-point the results are aggregated using a weighted prediction score. For a configuration c, the prediction score writes:

$$\hat{P}(y_{(\text{main,change-point})} = c|\text{nearby station}(ns)) = \frac{\sum_{ns} w_{ns} \mathbb{1}_{\{\widehat{\psi}(x_{ns})=c\}}}{\sum_{ns} w_{ns}}$$

where w_{ns} denotes the weight of the nearby station ns, and the final configuration is the one with the highest score

$$\hat{y}_{(\text{main,change-point})} = \arg\max_c \hat{P}(y(\text{main, change-point}) = c|\text{nearby station})$$

We compared two different weightings:

— inverse distance weighting: $w_{ns} = 1/d_{ns}$, where d_{ns} is the distance between the nearby ns and the main station,

— weighting proportional to the joint probability given in Table 4.1: $w_{ns} = p_c$.

In the probability-based weighting, when the highest score is reached by two different configurations (e.g. $c = 1$ and 10), the one with the shortest distance is selected.

Figure 4.8 presents the distribution of predicted configurations, after aggregation, for four variants: (a) balanced sampling with $\alpha = 5\%$, aggregated with distance; (b) balanced sampling with $\alpha = 1\%$, aggregated with distance; (c) imbalanced sampling with $\alpha = 5\%$, aggregated with distance; and (d) balanced sampling with $\alpha = 5\%$, aggregated according to prior probability. Across all figures, four predominant groups emerge consistently: G

(c=1 and 15), (G, E, E') (c=31 and 35), (E, E') (c = 10 and 23), and E (c = 8 and 22). Remarkably, these configurations correspond to the highest joint probabilities, p, as indicated in Table 4.1: G and (E, E') with p = 0.18, (G, E, E') with p = 0.04, and E with p = 0.01. This demonstrates that the classifier actually predicts the configurations which we believe are the most likely in the real data, even when these probabilities are not directly used in the procedure such as in variants (a) and (b).

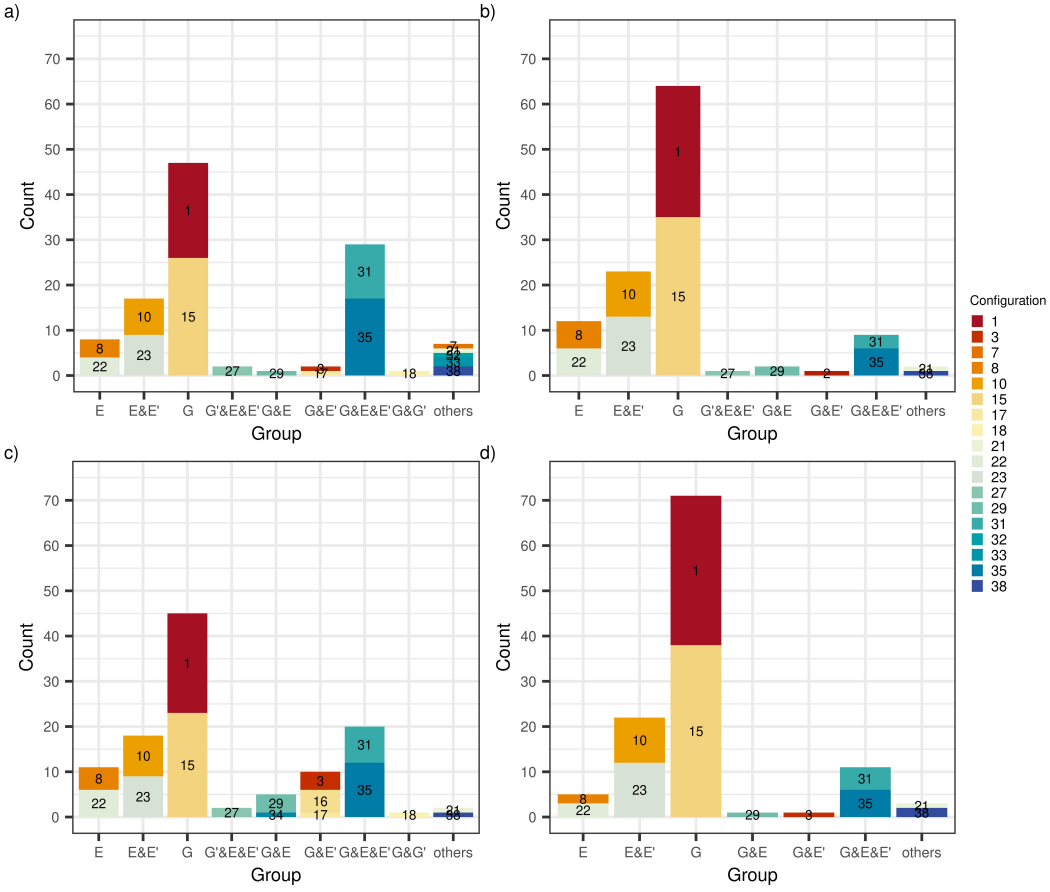


Figure 4.8 – Distribution of the final predicted configurations, after aggregation, from the real data with the Random Forest method. The numbers in color bars refer to the configuration number c among the 38 cases displayed in RSD Table (Table 4.1). Results are plotted for four cases: a) balanced sample learning with α = 0.05 and aggregated by distance, b) balanced sample with α = 0.01 and aggregated by distance, c) imbalanced sample with α = 0.05 and aggregated by distance, and d) balanced sample with α = 0.05 and aggregated by prior probability.

In variant (a), 47 of the change-points (i.e. 41%) are attributed to group G and 29 (i.e. 25%) to group (G, E, E'), after the aggregation. Analysis of the six test results before and after the prediction helps to understand the relatively high frequency of these two groups. In general, the test results can be of two sorts: either the six results correspond to a configuration in the Table 4.1, and in this case the predictive rule predicts the same result (as expected), or the result is initially not in the Table, and the predictive rule will select a configuration that is "close" to the initial configuration. Among all the test results going to group G, i.e. (1,1,1,0,0,0) for c = 1 and (-1,-1,-1,0,0,0) for c=15, about 75% are initially in the Table. This high percentage is consistent with the observation that many jumps are significant in the first three tests and insignificant in the last three, as seen

in Figures 4.6 and 4.7. The 25% of cases which are not initially in the Table differ from these configurations by one or two elements, e.g. case (1,0,1,0,0,0) differs from c=1 by only the 2nd element (the G-G' test). This case is then attributed to c=1 by the predictive rule when the absolute value of the t-value of the estimated jump in the G-G' series is close to the critical value, $\tau_{\alpha/2}$=1.96, in combination with smaller t-values in E-E', G'-E', and G'-E. For group (G, E, E'), the percentage of cases that are not in the Table is slightly more than 50%. Almost all these cases are either (1,1,1,0,1,0) or (-1,-1,-1,0,-1,0), which differ only by the 6th element (the G'-E test) from the final configurations c=31 (1,1,1,0,1,1) or c=35 (-1,-1,-1,0,-1,-1), respectively. Contrary to the G-G' series, the G'-E series has smaller t-values on average, hence the frequent 0 in the initial test results. The fact that almost all these cases are finally predicted as c=31 or 35 can be explained by the simultaneous occurrence of: high t-values in G-G' and G-E', a small t-values in E-E', a high t-values in G'-E', and a value close to the critical value for G'-E. An example is provided in Figure 4.9. In this example, one may also suspect that the t-value of G'-E' is excessively large, given the small value of the corresponding jump (-0.12) compared to the jumps in G-E, G-G', and G-E'. One way to reduce the occurrence of excessively high t-values in G'-E' is to increase the critical value of the test. For example if we set $\tau_{\alpha/2}$=2.58 ($\alpha = 0.01$), the test result here becomes 0 and the initial configuration becomes $c = 15$, which has a higher probability in Table 4.1 and is thus preferred.

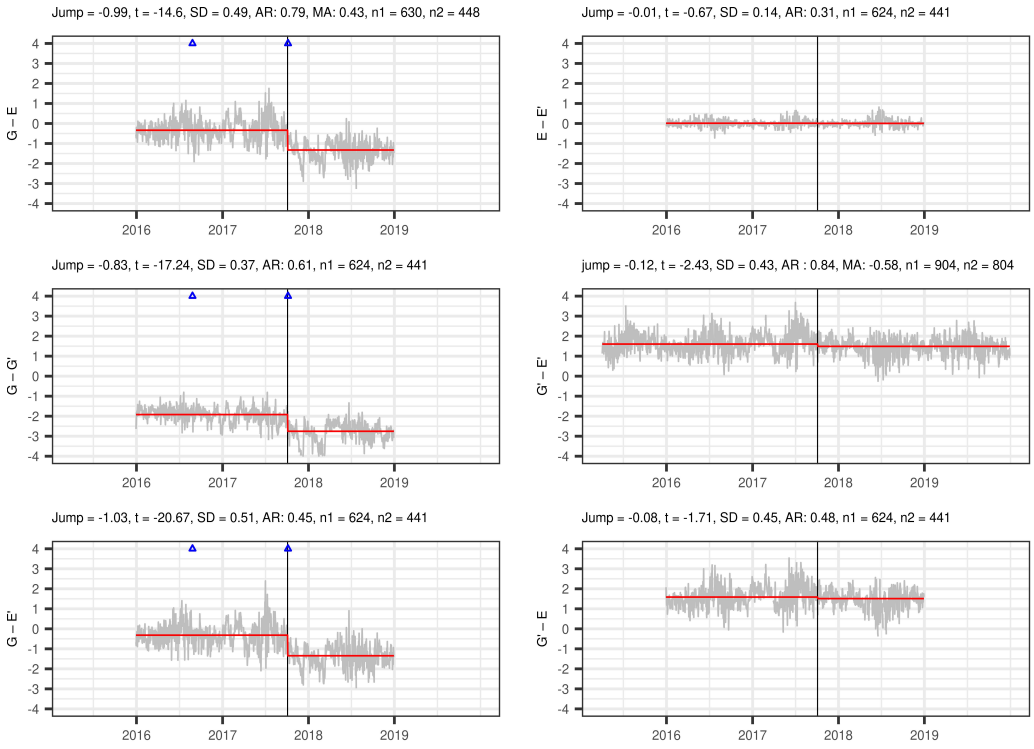


Figure 4.9 – Example of test result for station FAIR (Fairbanks, Alaska) with nearby station CLGO at a distance of 21 km. The series of IWV differences are shown in gray. The black vertical solid line shows the change-point detected in G-E by the segmentation (04 October 2017). The blue triangles indicate known equipment changes in the main station from the GNSS metadata. The horizontal red lines show the means estimated by the FGLS regression on the left and the right of the change-point in each series.

The impact of using $\alpha = 0.01$ is further illustrated on all tests with Figure 4.8b. Only 9 change-points are now assigned to group (G, E, E'), which is considerably smaller than with $\alpha = 0.05$. Actually, 10 change-points

moved to group G and 9 to group (E, E'). This difference can be understood by inspecting the distribution of t-values with respect to the corresponding critical values (2.58 vs. 1.96). Figure 4.6 shows that many t-values for E-E', G'-E, and G'-E' are smaller than 2.58 in absolute value. When these tests become insignificant, while the other three stay significant, the predicted configuration becomes c=1 or 15, and when G'-E' remains significant or close to 2.58, the predicted configuration becomes c=31 or 35. Additionally, many configurations with low probabilities (c=3, 7, 17, 18, 32, 33) have also disappeared.

Figure 4.8c shows the impact of the (probability-based) imbalanced sampling in the learning procedure. Overall, the results for the main groups are not much different compared to the balanced sampling. Two noticeable differences emerge, however. Firstly, group (G, E, E') reduces only slightly in size, from 29 to 20. The smallness of the impact is explained by the fact almost half of test results are initially in the Table 4.1 and are not changed by the prediction. Secondly, almost all the configurations with the lowest probabilities, such as c = 7, 21, 32, 33, 38, with $p \leq 5.6 \times 10^{-4}$, are removed. Other configurations, with slightly higher probabilities, but still with $p \leq 0.01$, such as c = 3, 16, 17 (G, E') and c = 29 and 34 (G, E), emerge or are reinforced, which is not wanted.

Figure 4.8d shows the variant (d) where the aggregation is based on the prior probabilities. The distribution is quite different from that based on distance (Figure 4.8a): more change-points are attributed to the preferred groups, 62% in G and 19% in (E, E'), fewer to other groups such as (G, E, E') and E, and many configurations of low probability disappear. The distribution is actually quite similar to that of variant (b), but in contrast to the latter, this variant keeps a high power in the test (thanks to α=0.05). As a result, with variant (d), group E is much smaller than with variant (b). In this respect, variant (d) is preferred among all four variants. Note, however, that there is a limitation in the usage of the aggregation procedure which holds for all variants: when there is only one nearby station (30 % of the cases), the aggregation has no impact and the final result is the one selected by the prediction rule. In variant (d), this explains why there are still configurations with a low probability (c=8/22, 3, 31/35, 21, 38).

4.1.6 Conclusions and perspectives

We proposed a post-processing method for the attribution of change-points detected by a segmentation scheme involving multiple series of differences (target minus reference). In our application, the each of the stations provides a GNSS series (G) and a reanalysis series (E). The segmentation is run on the G-E series and the goal of the attribution method is to predict if the inhomogeneity (jump in the mean) is in G or E. The method proceeds along the following steps:

1. **Data selection and pre-processing.** For each detected change-point in a main station (hereafter, the "main change-point"): (a) select nearby stations with a horizontal distance smaller than 200 km and height difference smaller than 500 m; (b) run the segmentation method on the G'-E' for each nearby data and select only homogeneous segments from the nearby to compare with the main; (c) correct the nearby series, G' and E', for the height difference with respect to the main station, so that all four series (G, E, G', and E') are representative of the same height; (d) form the six series of differences (G-E, G-G', etc.) and remove the outliers.
2. **Test the significance of the jumps.** For each main change-point and each of the six series: (a) identify the noise model; (b) fit a regression model including a jump at the position of the main change-point

when at least $n = 200$ consecutive points are available on the left and right of the change-point, using an iterative FGLS procedure; (c) test the significance of the jump at the significance level $\alpha = 0.05$.

3. **Use a predictive rule to predict the configuration.** For each nearby, the learned classifier will predict the configuration (i.e. which of the G, E, G', and E' series have a significant jump) corresponding to the six test results. When several nearby series are used, a weighted prediction score is computed to select the final configuration.

The method has been applied to a real data set of 494 cases (114 change-points from 49 main stations compared with 312 nearby stations). The data characterization showed that the data have a strong heteroscedasticity, with mean annual seasonal variation in the standard deviation around 70% (half-range), and a moderate autocorrelation, with a typical lag-1 correlation coefficient of 0.3. A FGLS test procedure was implemented to ensure an accurate inference. The predictive rule has been trained on the real data. Several classifiers have been compared for the predictive rule and the Random Forest was selected.

To our knowledge, both the FGLS regression test approach and the Random Forest classifier have never been used in the context of climate series homogenization.

The FGLS tests and the classification results of the studied data set have been assessed using i) our expertise of the data set (formulated out in two probabilistic rules) and ii) metadata informing about known equipment changes at the GNSS sites. Very consistent and plausible results were found from both the FGLS tests and the classification. With a significance level of 5% and employing a balanced sample for the learning step in the predictive rule, as well as aggregating results from nearby sources based on prior probability, the findings clearly indicate predominance of significant jumps in the series involving G (62%), (E, E') (19%), and (G, E, E') (10%) as expected. The remaining 9% of unexpected results are thought be linked with low detection power of the FGLS test when the noise is large (e.g. due to large distance between the main station and the nearby) and possibly random errors in the classification due to the smallness of the learning sample.

Some possibilities to further improve the method are: i) to use a bigger data set to improve the predictive method, ii) to refine the nearby selection rules to improve the robustness and the power of the test procedure (e.g., select nearby series with smaller percentage of gaps, shorten the distance between the main station and nearby), iii) compute the critical value used in the FGLS test from a more realistic empirical distribution. These options will be tested in a future work.

4.1.7 Appendix

4.1.7.1 Test table

Table 4.1 shows the theoretical test results one would obtain for the six paired difference series (G-E, G'-E', G-G', G-E', G'-E, E-E') with a perfect test method, for 54 different combinations of jumps in the four base series (G, E, G', E'). The jumps in the base series are coded on three values: 0 (no jump), +1 (upward jump), and -1 (downward jump). Here, only cases where either G or E, or both, have a jump are considered, because we always start we a change-point detected by the segmentation method in the G-E series (so either in G, in E, or in both). The corresponding results in the SD Table are thus theoretically coded on five different values (0,

1, 2, -1, and -2). However, in practice, a test result will be either reject (-1 or +1, where the sign indicates if the jump is upward or downward) or fail to reject (0) the null hypothesis of no jump. The RSD Table shows the corresponding practical results where the results are coded on three values only. The SD Table contains actually 46 unique combinations of the six test results and 8 duplicates highlighted by a colored background, while the RSD Table contains only 38 unique combinations. The duplicates are sorted out depending on probabilities (see Appendix B), i.e. those with lower probabilities are not counted.

4.1.7.2 Computation of prior probabilities

This section explains how we compute the conditional and joint probabilities reported in the Table 4.1. The notation hereafter is A (italics) for events and A (straight) for time series.

Let G, E, G', and E' represent the event "there is a jump" in each of the four base series, with the possible outcomes: 0=no jump, -1=the jump is downward, +1=the jump is upward. Let $P(A|B)$ represent the conditional probability of A given B. Let us assume that the events G and E are independent, and that G' and E' are independent. Now we reformulate the two rules that we stated in the Introduction and associate them with probabilities:

(R1) $P(G' \neq 0 | G \neq 0 \cup E \neq 0) = 0.1$.

(R2) $P(E = E') = 0.9$.

The values of 0.1 and 0.9 are chosen in a way to reflect the contrast between the corresponding events. Note that the first rule is stated more generally than in the Introduction as here it is now also conditional on a jump in series E. Indeed, in practice, it is unlikely that a jump in series G' occurs at the same time as in any other series. The first probability accounts thus for two incompatible events: $(G' \neq 0 | G \neq 0, E = 0)$ and $(G' \neq 0 | G = 0, E \neq 0)$, which we can suppose to be of similar probability $p_1 = 0.05$. It follows from the first rule that $P(G' = 0 | G \neq 0 \cup E \neq 0) = 0.9$. It follows from the second rule that $P(E' \neq E) = 0.1$, which also corresponds to two incompatible events $(E' = 0 | E = -1)$ and $(E' = 0 | E = +1)$, the probability of which we will also assume to be equal to $p_2 = 0.05$.

The conditional probabilities $P(G', E' | G, E)$ are obtained from the products of the individual probabilities $P(G', E' | G, E) = P(G' | G, E) \times P(E' | G, E)$. These probabilities are sufficient to distinguish the duplicates in the first 36 rows of the SD Table and the RSD Table (e.g. rows 1 and 18, in the numbering going from 1 to 54, have the same 6 values but are associated with different conditional probabilities). However, more duplicates appear in the RSD Table in the last 18 rows (row numbers 37 to 54). In order to distinguish them, we introduce prior probabilities associated to the events G and E. There are six different combinations: $(G, E) \in \{(+1, 0), (0, -1), (-1, 0), (0, -1), (+1, -1), (-1, +1)\}$. In the first 4 combinations, a single change occurs in either G or E, while in the latter 2, two changes occur simultaneously. The former is more likely, so we attribute it a probability of $p_3 = 0.225$, from which it results that the latter has a probability of $p_4 = 0.05$ which is deduced from the equation $4 \times p_3 + 2 \times p_4 = 1$.

Finally, the joint probabilities reported in Table 4.1 are obtained from the product of the conditional and the prior probabilities: $P(G, E, G', E') = P(G', E' | G, E) \times P(G, E)$.

4.2 Additional studies

4.2.1 Assessment of the stochastic model identification and parameter estimation

In the work presented in the previous section, we used the `auto.arima` function of the R package forecast (Hyndman and Khandakar, 2008) to identify the noise model in the time series of IWV differences and the arima function to estimate the model parameters. Results are presented in Figure 4.3 of the paper, for all the change-points tested from the main-nearby pairs, for each of the six series of differences.

The utilization of `auto.arima` in time series forecasting is widely acknowledged in the literature (Shaub, 2020; Yan et al., 2022; Radke et al., 2020; Lai and Dzombak, 2020). However, its skill for model identification has not been extensively discussed. In this subsection, we specifically evaluate its effectiveness within our application framework, i.e. the identification of four particular stochastic models: white noise (WN), AR(1), MA(1), and ARMA(1,1). We recall that, in a preliminary investigation, we tested ARMA(p, q) models up to order $p, q = 2$, but we found that only a negligible proportion of the time series were identified with order 2 and few of them had actually significant coefficients ($\alpha = 0.05$), so we limited ourselves to order $p = 1$ and $q = 1$.

Our examination below investigates the impact of the length of the series, N, and the coefficient values, ϕ and θ, on the variance of the coefficient estimates. We have seen from equations (2.20 to 2.23), introduced in subsection 2.2.3.8, that the variance of the coefficient estimates increases when the coefficients values decrease. To which extent the coefficient values impact the model identification performance is, however, not well known. The subsections below first assess the model identification performance of `auto.arima` in terms of the probability of identifying the true model. Then, the uncertainty of the coefficients estimated with the `arima` function is examined assuming the model identified is correct. We also investigate the influence of gaps in the time series on both model identification and parameter estimation quality. This is a situation frequently encountered with real data. Finally, we discuss the possible impact of the uncertainty in the model identification and parameter estimation in the case of the real data presented in the paper.

4.2.1.1 Simulation set-up

The simulations involve varying the length of time series, ranging from $N = 200$ to 2000 data points. Coefficients for the AR(1) and MA(1) models are chosen from the range of 0.1 to 0.6, covering typical values encountered in real data. For the ARMA(1,1) model, the study examines six cases where the sum $s = \phi + \theta$ goes from 0.1 to 0.6, where ϕ is varied from 0.8 to -0.8 and θ is adjusted accordingly. Importantly, cases where $\phi = 0$ or $\theta = 0$, representing a MA(1) or an AR(1) model, respectively, are excluded from the ARMA(1,1) discussion. The analysis comprises 1000 simulation runs and the variance of the white noise (or innovation) is set to 1.

4.2.1.2 Noise model identification results

The relationship between series length and the efficiency of model identification is presented in Figure 4.10 and Figure 4.11. All the plots demonstrate a consistent trend: the True Positive Rate (TPR), which quantifies the fraction of time series for which the true model is correctly identified, improves as the series length increases. However, the slope of the TPR curves varies depending on the model type and coefficient values. Three interesting features can be mentioned. Firstly, the dependency of TPR on series length for the AR(1)

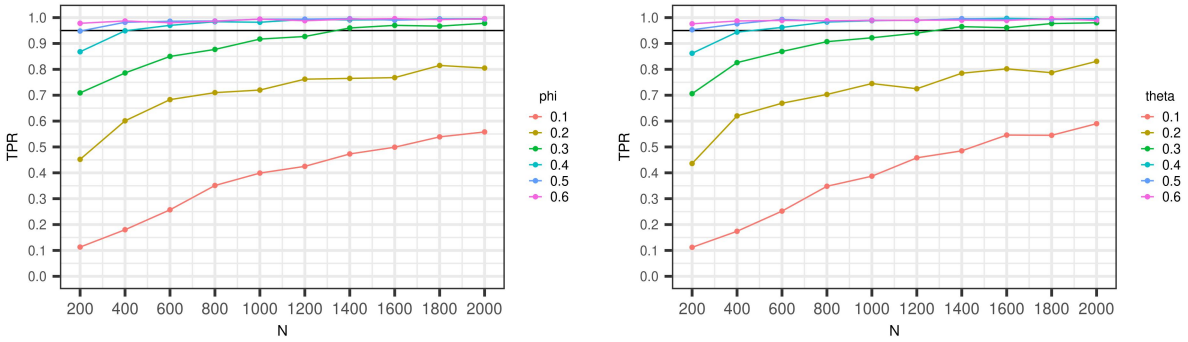
and MA(1) models appears to be strongly resembling for similar coefficient values. The slope of the TPR curves is gentler for larger coefficient values and becomes steeper as the coefficient magnitude reduces. This behavior is certainly in connection with the impact of coefficient values on estimation uncertainty (smaller coefficients leading to larger uncertainty, discussed in Chapter 2.2.3). Notably, for coefficient values of 0.3 or larger, the TPR exceeds 70% when $N = 200$ and 95% when $N = 1400$, for both models. However, with smaller coefficients, a TPR value of 95% is never reached, even for a series length of up to 2000. Secondly, for the ARMA(1,1) model, the shape of the TPR curve is clearly dependent on the sum of the coefficients, $s = \phi + \theta$, where inverting the values of ϕ and θ yields similar TPR results. It is also seen that the slope of the TPR curves if steeper when either the absolute values of coefficients are larger or when the sum is larger. For example, for a scenario wherein the sum of the two coefficients is $s = 0.3$, the TPR goes from roughly 40% for $N = 200$ to 95% for $N = 800$ when $\phi = 0.7$ and $\theta = -0.4$ (and similarly when $\phi = -0.4$ and $\theta = 0.7$). However, with the combination $\phi = 0.6$ and $\theta = -0.3$, the TPR of 95% is only achieved for a longer series $N = 1800$. Lastly, several cases exist where the TPR remains notably low even with extended series lengths, suggesting inherent challenges in model identification. For example, when the sum of the coefficient values is 0.3, coefficient pairs such as $(0.1, 0.2)$, $(0.2, 0.1)$, $(-0.1, 0.4)$, and $(0.4, -0.1)$ result in TPRs below 10%, even when N reaches 2000.

In Figures 4.12 and 4.13, the same results are plotted in an alternative way to enhance the influence of model coefficients on the efficiency of the model identification.

For the AR(1) and MA(1) models shown in Figure 4.12, the TPR ascends as coefficients increase, reaching a value near 1 for ϕ or θ values of approximately 0.6 for all series lengths. For smaller coefficients, the TPR displays greater dispersion across series lengths. It is interesting to examine which alternative model is actually identified when TPR is smaller than 100%. Figure 4.14 shows that a negligible fraction of cases are identified as ARMA(1,1) when the true series is AR(1) or MA(1). This can be easily understood by the parsimony effect embedded in the BIC model selection. However, when the coefficient of the AR(1) model becomes small, a non negligible fraction of cases are identified as MA(1) and vice-versa. In the extreme case when the coefficients are close to zero, they may be identified as WN (e.g., about 30% when ϕ or θ=0.1).

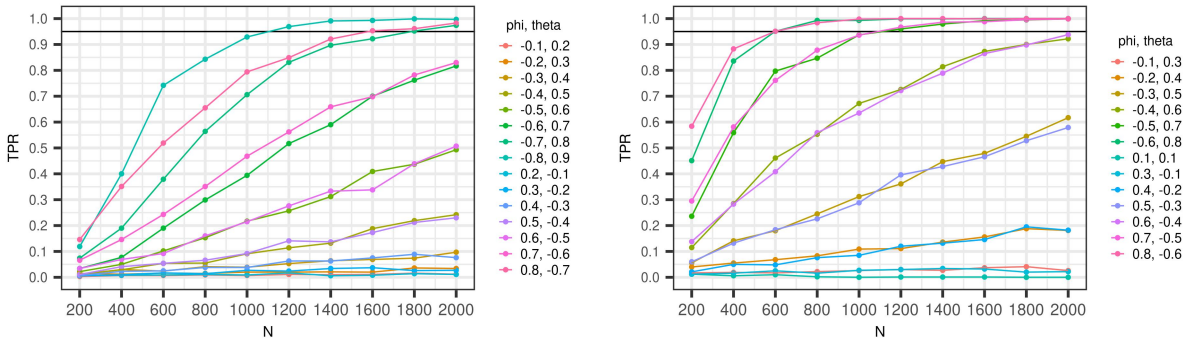
For the ARMA(1,1) model (Figure 4.13), the TPR behavior is very different depending on whether the sum of ϕ and θ is smaller or larger than 0.3, showing either single or dual dips in the TPR curve respectively. It is noteworthy that these dips consistently appear when one of the coefficients nears zero. For example, with $s = 0.3$, when either $\phi = 0.1$ and $\theta = 0.2$ or its converse, the TPR goes to zero. Conversely, for $s > 0.3$, e.g. when $s = 0.5$, dual dips emerge for combinations $\phi = 0.4$ and $\theta = 0.1$, and $\phi = 0.1$ and $\theta = 0.4$. The reason why the TPR is so low in this case can be understood by examining the details of the model identification outcomes shown in Figure 4.15. Indeed, in situations where one parameter is small and one is large, it is logical that a simpler model is preferred, i.e. a AR(1) or a MA(1). This is clearly observed on this Figure: an ARMA(1,1) is correctly identified only when both ϕ and θ are not close to 0, whereas it is predominantly identified as AR(1) when θ is small and as MA(1) when ϕ is small. Moreover, the fraction of correct identification not only depends on the values of ϕ and θ but also on their sum, $s = \phi + \theta$. Notably, Figure 4.15e shows that for the latter example with $s = 0.5$, over 85% of the series are classified under the AR(1) model (note that this simulation is specific to the case $N = 1000$). However, a special situation is observed in Figure 4.15a for $s = 0.1$, where a

significant fraction of cases are identified as WN even when ϕ and θ get large. Indeed, when s is close to zero, the values of ϕ and θ are nearly opposite and the stochastic process is actually close to a white noise process. Identifying an ARMA model when $\phi = -\theta$ is an ill-posed problem which leads to infinite variance in both estimated parameters as reflected from Eqs. (2.22 and 2.23). This issue is referred to as parameter redundancy or over-parameterization (Shumway and Stoffer, 2011).



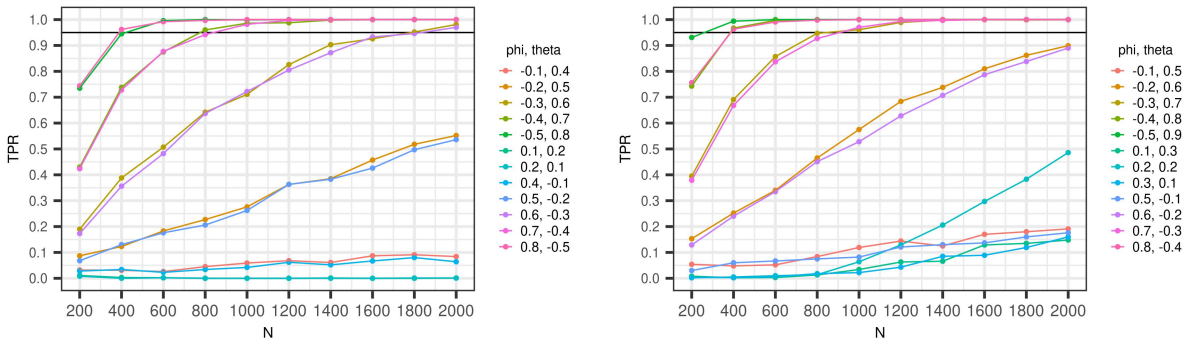
(a) AR(1) (b) MA(1)

Figure 4.10 – True Positive Rate (TPR) against series length for AR(1) model (left) and MA(1) model (right).



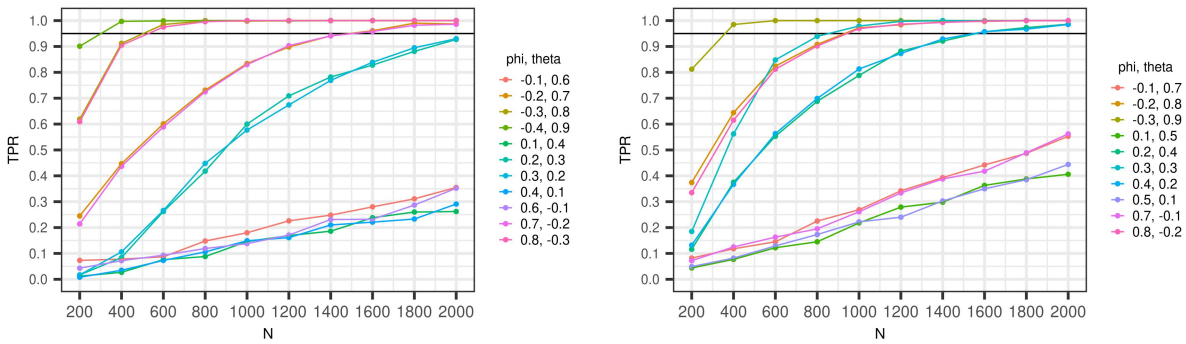
(a) s = 0.1

(b) s = 0.2



(c) s = 0.3

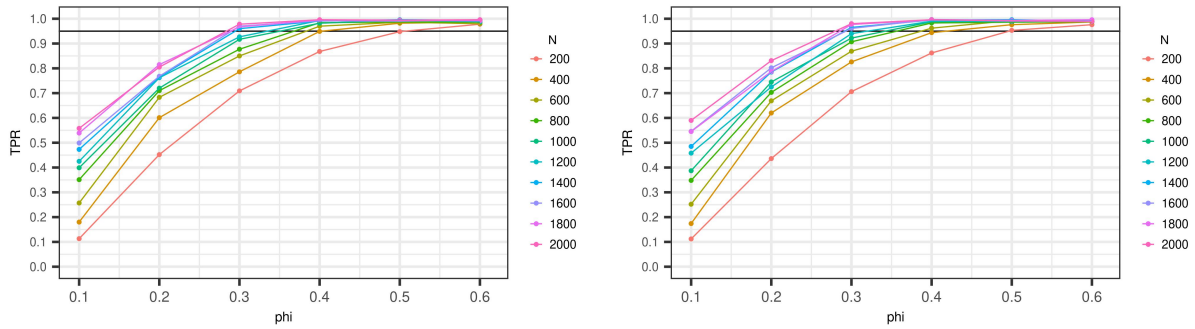
(d) s = 0.4



(e) s = 0.5

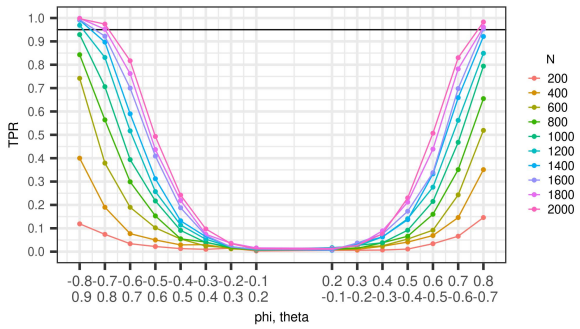
(f) s = 0.6

Figure 4.11 – True Positive Rate (TPR) against series length for ARMA(1,1) model. Plots (a) to (g) depict the results for six variations, with coefficient sums $s = \phi + \theta$ ranging from 0.1 to 0.6.

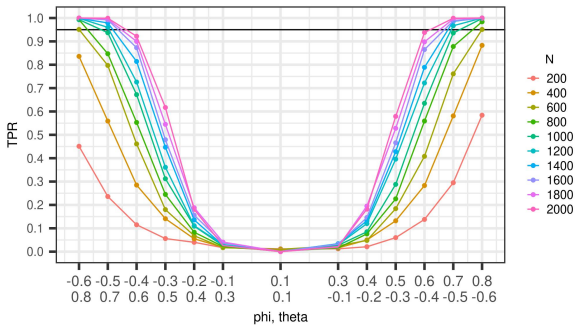


(a) AR(1) (b) MA(1)

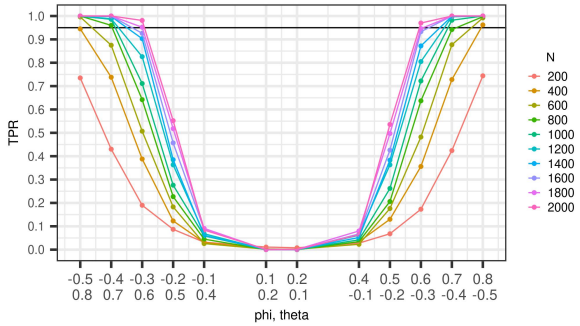
Figure 4.12 – True Positive Rate (TPR) as a function of the coefficient value for the AR(1) model (left) and the MA(1) model (right).



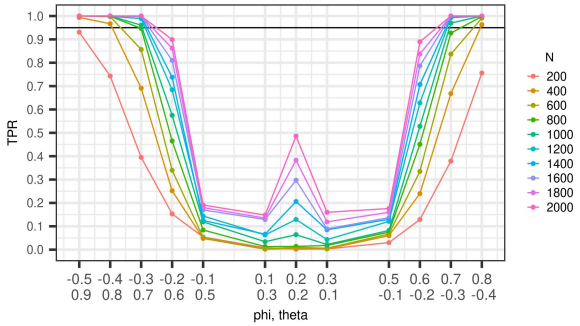
(a) s = 0.1



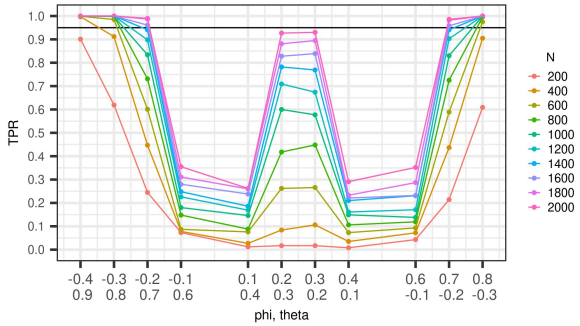
(b) s = 0.2



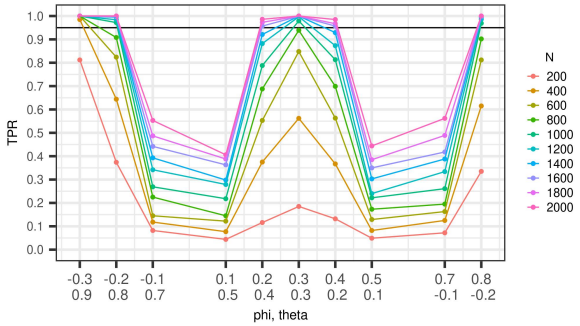
(c) s = 0.3



(d) s = 0.4

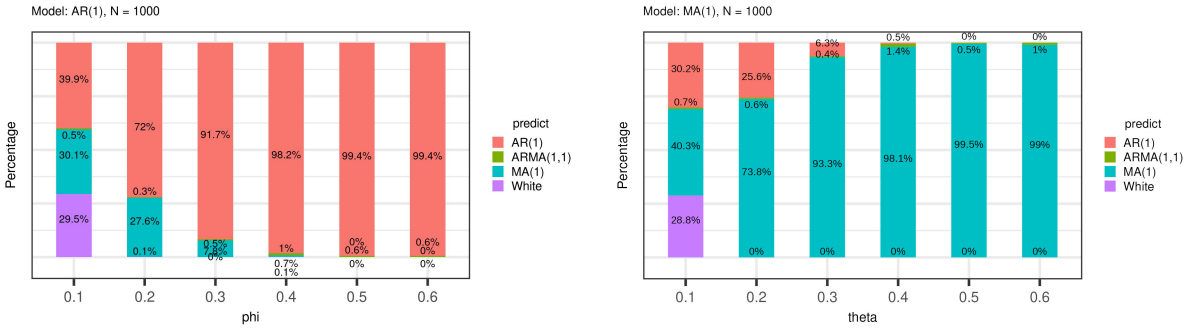


(e) s = 0.5



(f) s = 0.6

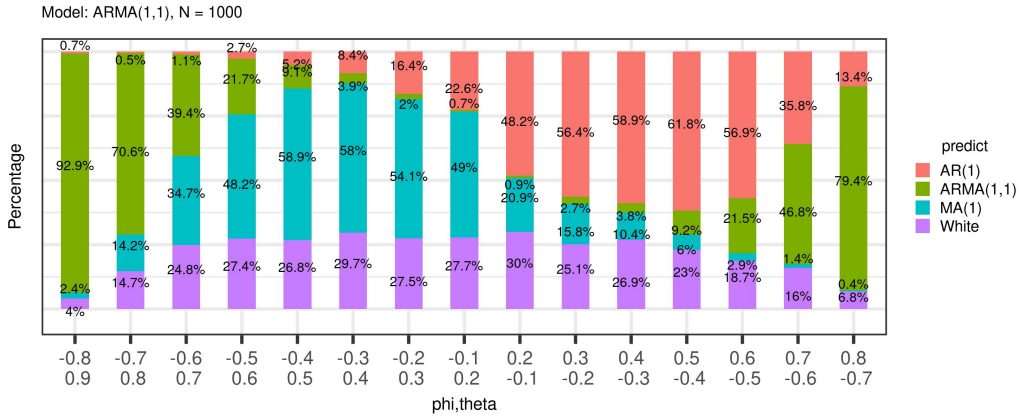
Figure 4.13 – True Positive Rate (TPR) as a function of the coefficient values for ARMA(1,1) model. Plots (a) to (f) depict the results for six variations, with coefficient sums $s = \phi + \theta$ ranging from 0.1 to 0.6.



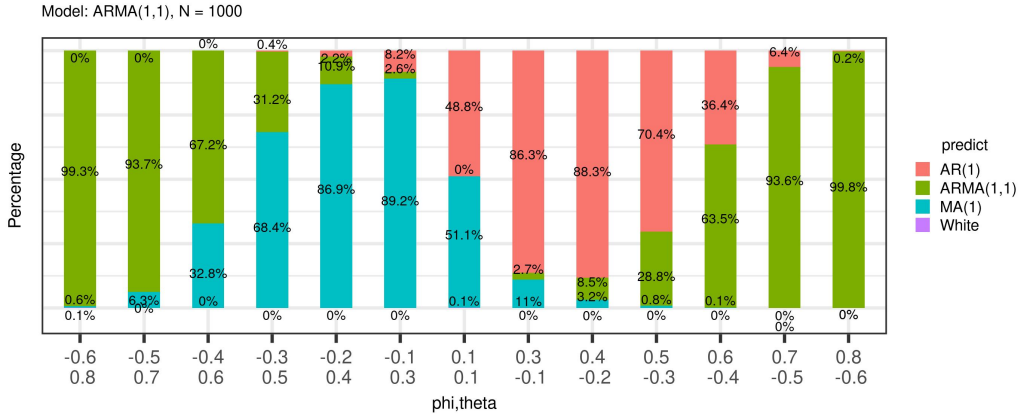
(a) AR(1)

(b) MA(1)

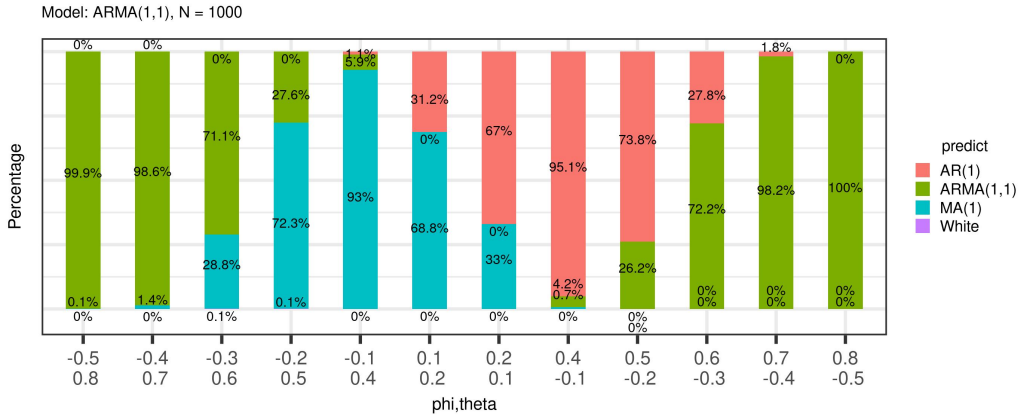
Figure 4.14 – Outcomes from the model identification for AR(1) (left) and MA(1) (right) as a function of coefficient values, specifically for series lengths of 1000.



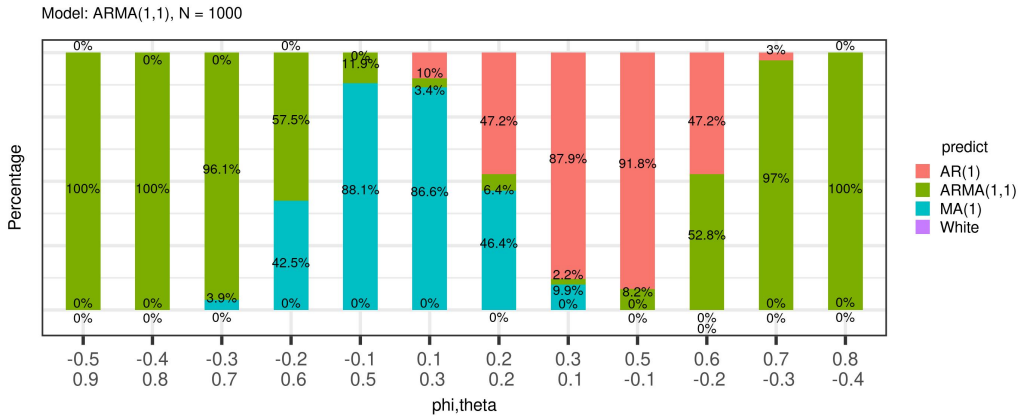
(a) s = 0.1

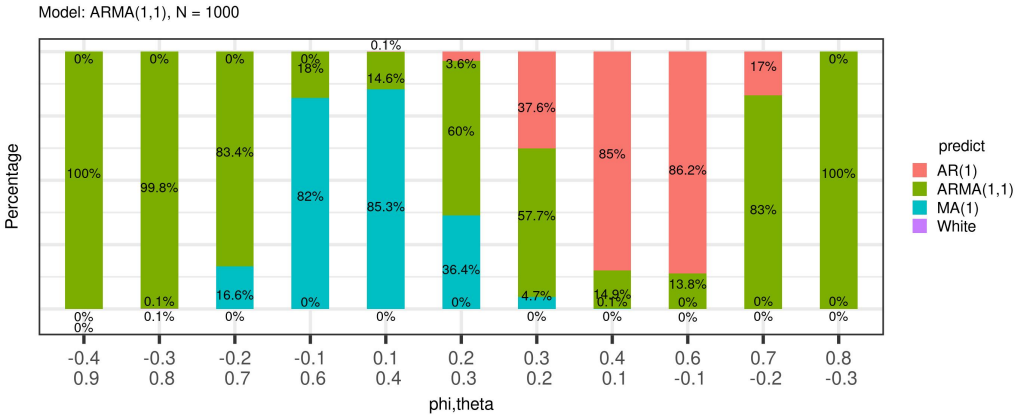


(b) s = 0.2

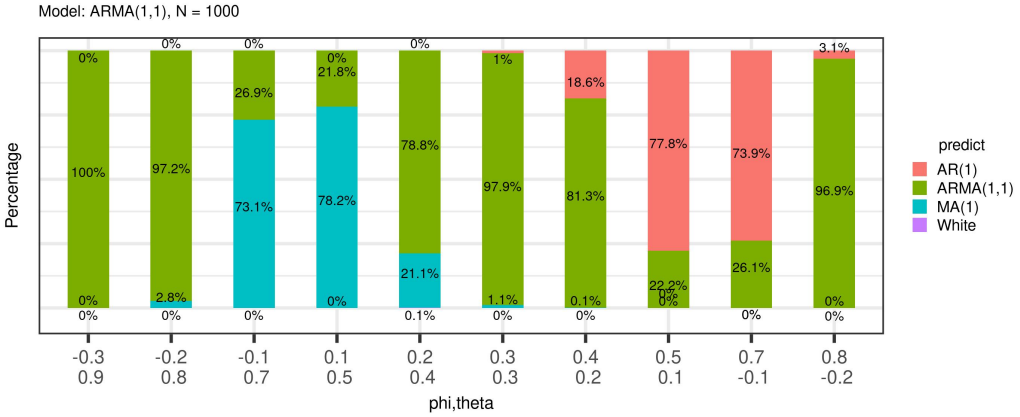


(c) s = 0.3





(e) s = 0.5

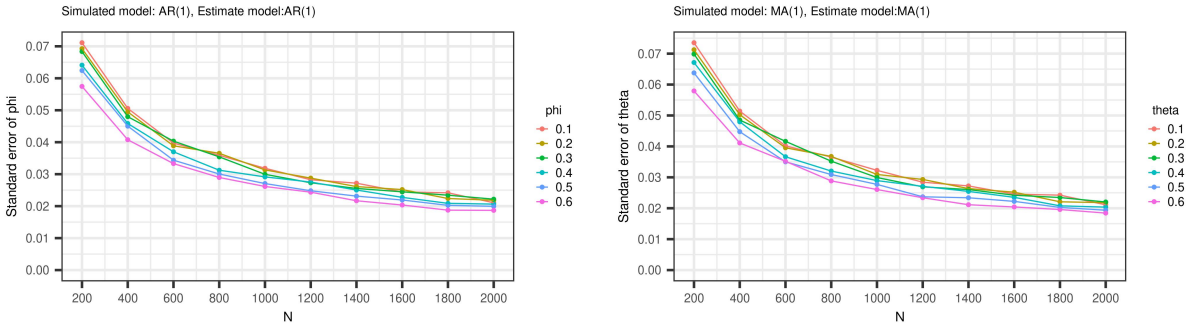


(f) s = 0.6

Figure 4.15 – Outcomes from the model identification of ARMA(1,1) as a function of coefficient values, specifically for series lengths of 1000. Plots (a) to (f) depict the results for six variations, with coefficient sums $s = \phi + \theta$ ranging from 0.1 to 0.6.

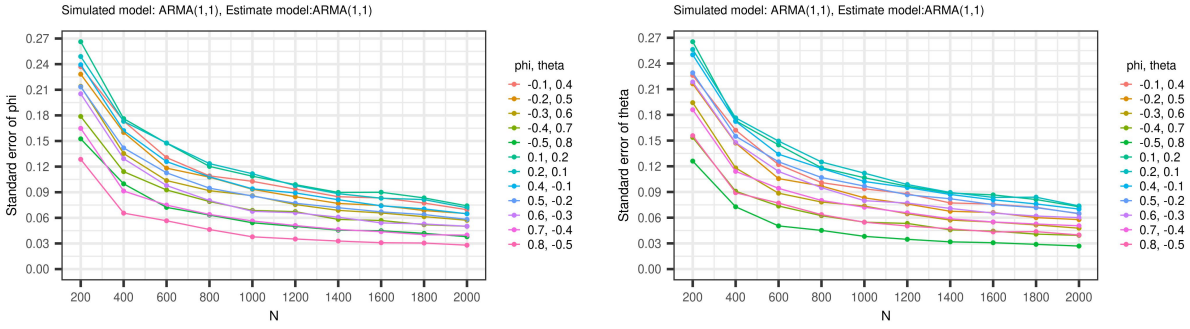
4.2.1.3 Parameter estimation results

Figures 4.16 and 4.17 display the mean standard error of the estimated coefficients for AR(1), MA(1), and ARMA(1,1) models obtained from the simulations. These values are in good agreement with the estimates computed from equations (2.20) to (2.23) presented in subsection 2.2.3. From these plots, it is convenient to predict the uncertainty in the estimated parameters when designing an experiment or analysis with real data. For example, for an assumed ARMA(1,1) process with $\phi = 0.6$ and $\theta = -0.3$, as reported in the paper, a standard error close to 0.21 is expected on both $\hat{\phi}$ and $\hat{\theta}$ when $N = 200$, which can be decreased to 0.08 for both estimates when $N = 1000$.



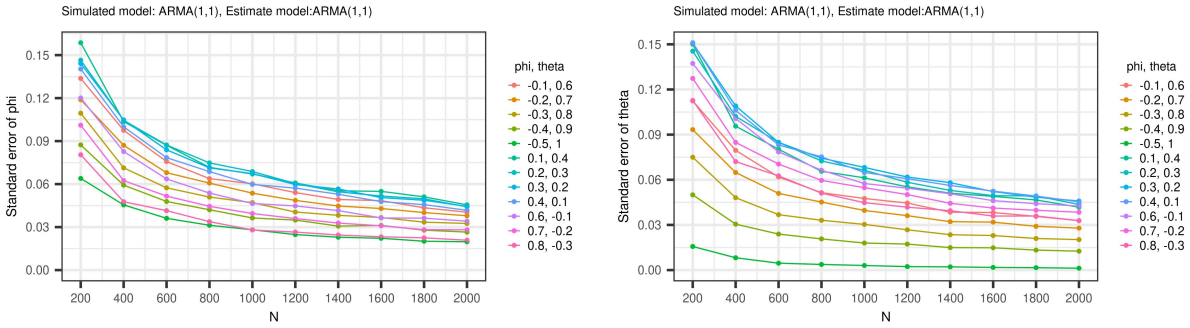
(a) AR(1) (b) MA(1)

Figure 4.16 – Mean coefficient standard error as a function of time series length for models: AR(1) with $\phi = 0.3$ on the left, MA(1) with $\theta = 0.3$ on the right.



(a) phi (b) theta

Figure 4.17 – Mean coefficient standard error as a function of time series length for the ARMA(1,1) model with $s = \phi + \theta = 0.3$.



(a) phi (b) theta

Figure 4.18 – Mean coefficient standard error as a function of time series length for the ARMA(1,1) model with $s = \phi + \theta = 0.5$.

4.2.1.4 Impact of gaps in the time series

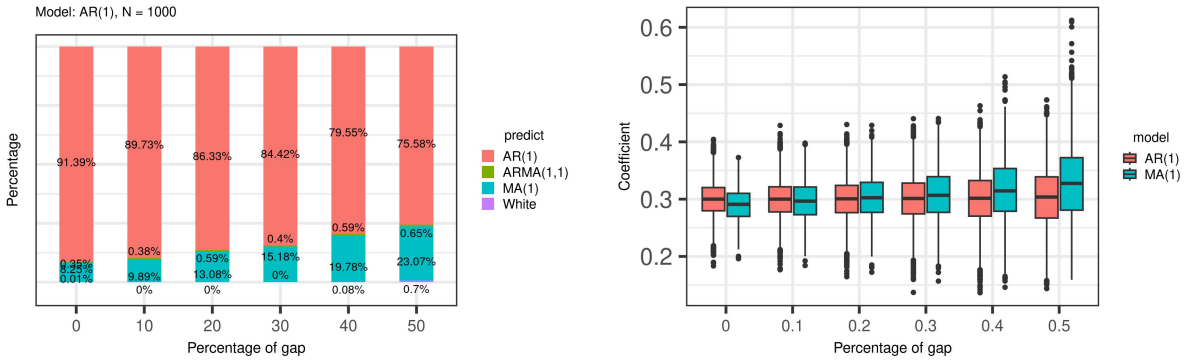
In practice, observational data series frequently exhibit measurement gaps. While some studies have investigated the impact of gaps on stochastic model parameter estimations, and have proposed modified

estimation procedures assuming a known model, the impact of gaps on model identification has not been much studied. Intuitively, one can expect that missing data will weaken the correlation structure, making it more challenging to correctly identify the model. The simulations below investigate the performance of the `auto.arima` function in the presence of gaps for the special case of an AR(1) process with a value of $\phi = 0.3$. In this simulation setup, the length of the time series remained constant, while a fraction of data points were replaced with NA values.

Figure 4.19 shows the percentage of identified models when the true model is an AR(1) (left) and the corresponding coefficient estimates (right) when the fraction of missing data varies between 0% and 50%.

The model identification results show that the AR(1) process is generally well identified, with a small fraction being MA(1). The TPR of the AR(1) drops from 91% to 76% when the fraction of gaps rises from 0 to 50%. Interestingly, even with half the data missing, the ability to detect a serial correlation in the series (either as AR(1), MA(1) or ARMA(1,1)) remains high at around 99%, while the rate of misclassifying AR(1) as white noise is under 1% (when $\phi = 0.3$).

Figure 4.19b presents the coefficient estimates obtained with the `arima` function, which takes missing data (in the form of NA values) into account. The plot displays the distribution of $\hat{\phi}$ and $\hat{\theta}$ when either an AR(1) or a MA(1) model is identified while the data are simulated from an AR(1) with ϕ=0.3. Firstly, for the correctly identified AR(1) cases, the estimates average consistently around 0.3; however, the scatter increases with increasing data gaps. Secondly, in cases where MA(1) is identified, the median of coefficients is about 0.33. The slightly larger coefficient value of the MA(1) process compared to the AR(1) process is easily explained by the fact that these processes have a similar value of the autocorrelation function (ACF) at lag 1 for these specific coefficient values.



(a) Model identification

(b) Coefficient estimates

Figure 4.19 – Percentage of identified models and the associated coefficient estimate against the percentage of missing data when simulating the AR(1) model with $\phi = 0.3$.

4.2.1.5 Interpretation of results from real data

The simulation results presented in the previous subsection can help to assess the reliability of the data characterization and to interpret the results presented in the paper.

First, we examine the uncertainty in the identification of the three main stochastic models with their typical (average) coefficients found in the paper. Recall that the minimum series length considered in the paper is N=400, composed of 200 points on each side of the change-point, and in practice, it often exceeds N=1000. Beginning with the AR(1) model with $\phi = 0.3$, we expect a correct identification of this model with a TPR=80% for a sample size of $N = 400$, and a TPR exceeding 90% when $N > 1000$ (green line in Figure 4.10a). For the MA(1) model with $\theta = 0.2$, TPR is about 60% for N=400 and rises to 83% for $N = 2000$ (olive line in Figure 4.10b). For the ARMA(1,1) model, we consider two coefficient combinations: (a) $\phi = 0.6$ and $\theta = -0.3$, and (b) $\phi = 0.3$ and $\theta = 0.2$. For the former, we expect a TPR of 40% at N=400 which rises to 95% by N=1800 (purple line in Figure 4.11c). For the latter, TPR initially lingers 10% when $N = 400$ but approaches 95% only at $N = 2000$ (cyan-green line in Figure 4.11e). The standard errors for these two ARMA(1,1) cases are std[$\hat{\phi}$]=0.069 and std[$\hat{\theta}$]=0.082 in case (a), and std[$\hat{\phi}$] ~ std[$\hat{\theta}$] ~ 0.065 in case (b), according to Equations (2.22) and (2.23).

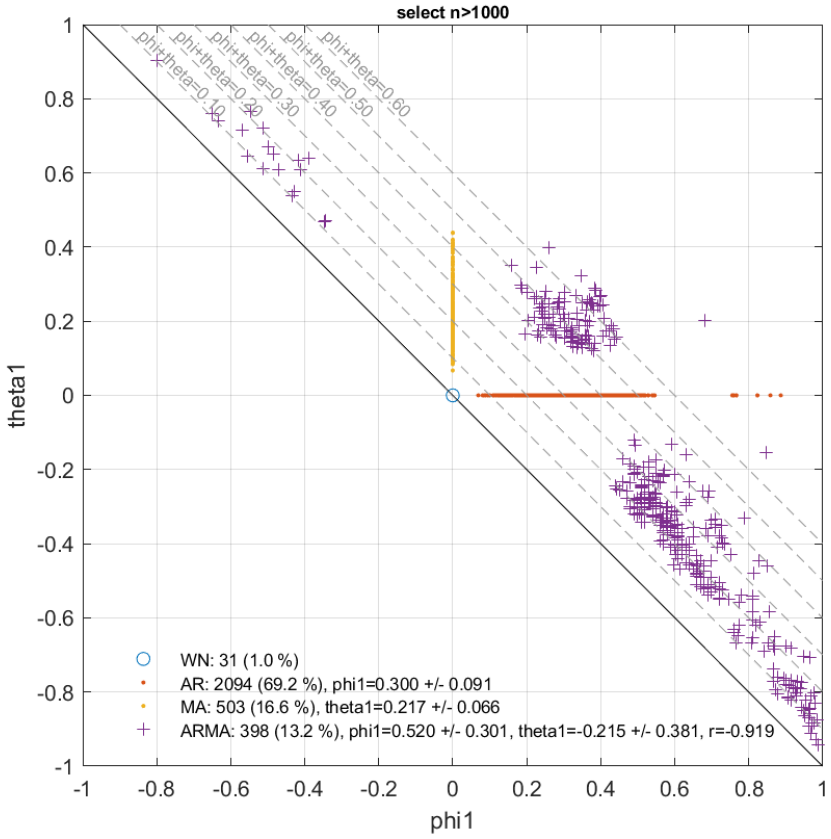


Figure 4.20 – Coefficient estimates, $\hat{\theta}$ vs. $\hat{\phi}$, for all six series of IWV differences and all (494) main-nearby pairs in the real dataset, with series length $N \geq 1000$. The colors correspond to different models (WN in blue, AR(1) in orange, MA(1) in yellow, and ARMA(1,1) in purple).

Next, we examine the dispersion of the estimated coefficients in the real data. Figure 4.20 shows the coefficient estimates, $\hat{\phi}$ and $\hat{\theta}$, from the data analysed in the paper, restricted to series with N>1000 data. This plot complements Figure 4.3 of the paper, but without separating the results for the data combinations. The

representation of the results in the ($\hat{\theta}$, $\hat{\phi}$) plane gives a special insight into the characteristics of the identified ARMA(1,1) models.

The distribution of the $\hat{\phi}$ and $\hat{\theta}$ coefficients for MA(1) and AR(1) models show values extending over a quite large range, from about 0.05 (for both coefficients) to 0.45 for $\hat{\theta}$ and 0.9 for $\hat{\phi}$. The TPR for the coefficient values below 0.1 is smaller than 40% (green curves in Figure 4.12a and b), i.e. more than 60% of these time series could be mis-identified as white noise (30%) or MA(1) instead of AR(1) and vice-versa (30%) as revealed by Figure 4.14a and b. For coefficient values above 0.3, the TPR exceeds 90%, so these case are correctly identified and estimated. The uncertainty in the coefficient estimates ($\sim$ 0.03, Figure 4.16) certainly adds to the scatter in the observed range of values.

As for the identified ARMA(1,1) models, the estimated coefficients exhibit a very large scatter, covering almost all possible values, from -1 to +1, except around zero. The ($\hat{\phi}$, $\hat{\theta}$) points appear actually to fall into two major and one minor groups, depending primarily on $\hat{s} = \hat{\phi} + \hat{\theta}$. the first major group aligns more or less on the $\hat{s} = 0.2$ line and includes $\hat{\phi}$ values in the range 0.4 to 0.99. Referring to Figure 4.15b for the case $s = 0.2$, the values of $\hat{\phi}$ at the lower end have very low probability to be identified as ARMA(1,1), which means that this class may be under-represented in our analysis. At the higher end, the values align more on the $\hat{s} < 0.1$ line, and these series, although identified as ARMA(1,1), are in reality close to a white noise. The standard error of the coefficients for the latter cases becomes very large, with makes their interpretation difficult. The few cases in the minor group where $\hat{\phi}$ takes negative values and $\hat{\theta}$ positive values are close to $\hat{s} \sim 0.1$, and may just be some examples of biased estimates from the first major group.

The second major group of the ARMA(1,1) results is centred on ($\hat{\phi}$, $\hat{\theta}$) = (0.3, 0.2). As was shown in Figure 4.3 of the paper, these cases correspond to all the E-E' series plus some other series when the distance between main and nearby is larger than 50 km. The scatter in this group is relatively small, which can be explained by the higher accuracy of the $\hat{\phi}$ and $\hat{\theta}$ estimates in this group compared to the previous (see 4.18). However, the probability of correct identification for this group remains only around 60% (Figure 4.15e), with almost 40% going to the AR(1) group. It is thus probable that the ARMA(1,1) cases from both groups are under-represented in our results, and that the observed AR(1) and MA(1) groups include the latter.

4.2.2 Understanding of test and prediction results

In this subsection, we complement the discussion on the test and prediction results given in Section 4.1.4.3 of the paper. First, our attention is on the results of the balanced sample learning (variant (a) of Figure 4.26. We will focus on cases where combinations of the six tests initially not in the Test Table (Table 4.1 presented in the paper). Second, we discuss the impact of the distance on the test results. Next, we analyse the distribution of the predicted configurations before aggregation and, last, we discuss the uncertainty in the final classifier.

Further discussion of the results of variant (a). Among the 494 tested cases, 225 correspond to configurations that are not in the Test Table. These cases can be assembled in different groups, of which the 3 main are presented in Table 4.5. Group 1 contains 52 cases, of which 32 correspond to test results of (-1,-1,-1,0,-1,0) and 20 of (1,1,1,0,1,0). Group 2 contains 37 cases, among which 22 correspond to test results of (-1,0,0,0,0,0)

and 15 of (1,0,0,0,0,0). Group 3 contains 14, where 10 cases correspond to test results of (-1,0,-1,0,0,0) and 4 of (1,0,1,0,0,0).

For the cases in group 1, the initial three tests have usually large t-values, while the subsequent three tests present lower, frequently insignificant t-values (which is a characteristic of configuration c=1). However, the cases in group 1 have a significant G'-E' test (this configuration is not in Table 4.5). This is due to the fact that the G' and E' series are co-located, resulting in a lower noise level compared to the other two series (E-E' and G'-E), which remain influenced by the distance (larger noise at larger distance). The reason why these cases are predicted into different configuration is discussed below.

The groups 2 and 3 are most frequently observed at greater distances, averaging around 114 km and 126 km, respectively. Group 2 becomes more prominent when the t-values in G-E' are insignificant, while Group 3 corresponds to cases when the t-values in G-E' are only slightly larger than the critical value (the average values of these t-values is 2.24 for the upward shift and -2.33 for the downward shift).

Examining the prediction outcomes for these three groups reveals two points. First, a general observation is that predictions typically adjust only one or two elements among the six. Second, the way the six tests relate to each other and to the critical value, determines which elements are adjusted. These underlying rationales is evident across all groups.

In the first group, with test results such as (-1,-1,-1,0,-1,0), configurations are predicted such as c =15 (where G'-E' goes from -1 to 0), c =16 (E-E' goes from 0 to -1), and c =35 (G'-E goes from 0 to -1). The clear predominance of prediction configuration c = 35 is further discussed in the subsection 4.1.5.3 of the paper. Analyzing further, the configuration predicted as c =15 has the following t-values for the 6 series: $(-7.5, -4.9, -7.5, -1.4, -2, -0.6)$. Here, the G'-E' is close to the critical value (1.96) but considerably lower than the first two tests (G-G' and G-E'), while G'-E is far to the critical value. Consequently, it is intuitive to shift G'-E' to being insignificant and predicting c = 15 instead of 35.

In group 2, five test results are insignificant. The classifier adjusts either the G-E' and G-G' result or the E-E' and G'-E' result. This leads to the predicted configurations 1, 15, 8, or 22, mainly because they are the only configurations with three zeros in the Test Table (Table 4.1).

Finally, the classifier predicts the cases of the group 3 to configurations 1 and 15, which differ from the test results only by the G-G' result. The average absolute t-values for G-G' in this group stands at 1.6, nearing the critical value (1.96) more than the G-E', E-E', and G'-E results which have average absolute t-values of 0.4, 0.7, and 0.8, respectively.

Group	Test results						Predicted							Count
	G-E	G-G'	G-E'	E-E'	G'-E'	G'-E	Config	G-E	G-G'	G-E'	E-E'	G'-E'	G'-E	
1	-1	-1	-1	0	-1	0	15	-1	-1	-1	0	0	0	1
							16	-1	-1	-1	-1	-1	0	1
							35	-1	-1	-1	0	-1	-1	30
	1	1	1	0	1	0	1	1	1	1	0	0	0	2
							31	1	1	1	0	1	1	18
2	-1	0	0	0	0	0	15	-1	-1	-1	0	0	0	17
							22	-1	0	0	1	0	-1	5
	1	0	0	0	0	0	1	1	1	1	0	0	0	2
							8	1	0	0	-1	0	1	13
3	-1	0	-1	0	0	0	15	-1	-1	-1	0	0	0	10
	1	0	1	0	0	0	1	1	1	1	0	0	0	4

Table 4.5 – (left part) Three main groups of test results which are not in the Table 4.1; (right part) configurations predicted by the classifier and number of cases (rightmost column).

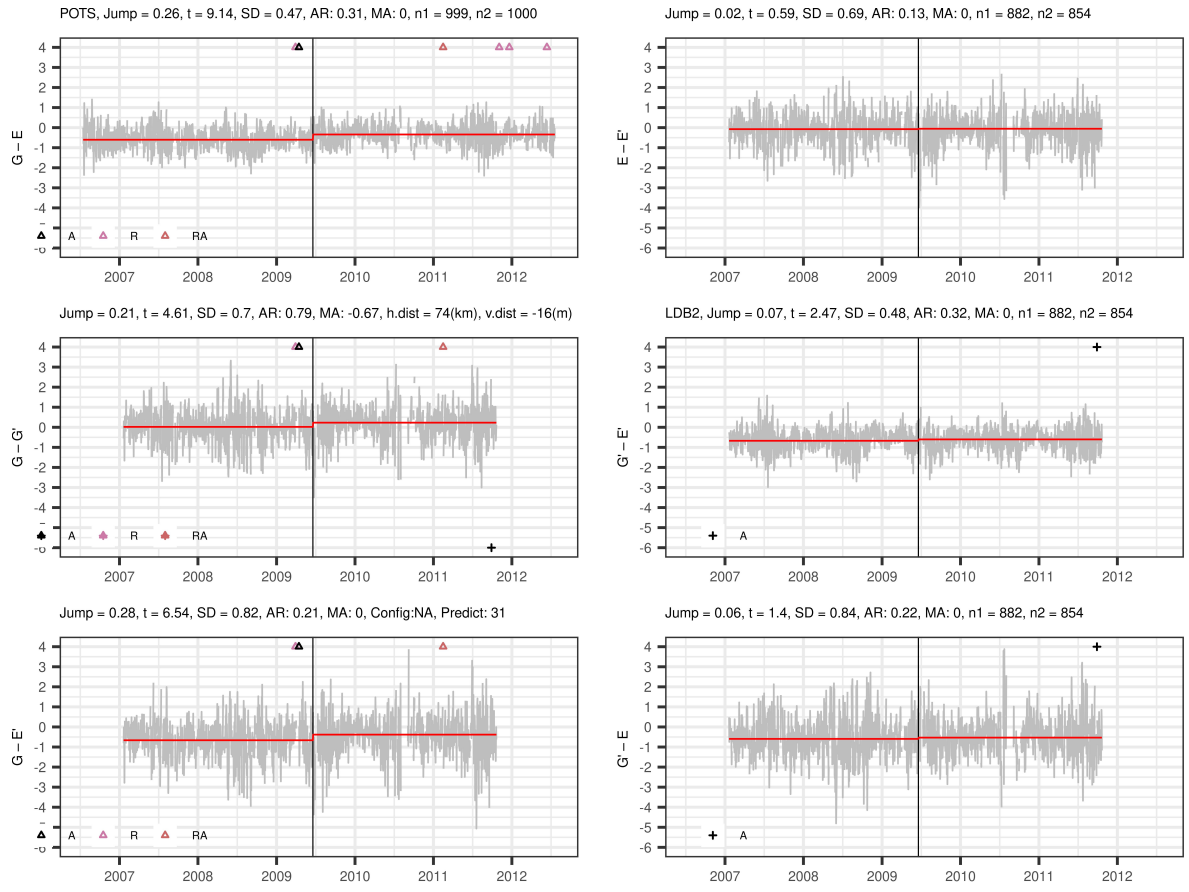


Figure 4.21 – Example of test results for station POTS (Postdam, Germany) with nearby station LDB at a distance of 74 km. The series of IWV differences are shown in gray. The black vertical solid line shows the change-point detected in G-E by the segmentation (19 June 2009).

Impact of distance on test results. In the subsection 4.1.4.3 of the paper, we have observed that the distance between the main and nearby stations has an impact on the standard error of the jump, and

therefore on the t-value. In particular, the standard error of estimated jumps is clearly smaller for short distance (lower than 50 km). Here, we discuss in more details this point according to the noise level of the series.

First, recall that the variance of the GLS jump estimator, $\hat{\delta}_{GLS}$, is given by Equation (4.6): $var[\hat{\beta}_{GLS}] = (X'\Sigma_0^{-1}X)^{-1}$, where Σ_0 is the variance-covariance matrix of the noise. In the simplest scenario of independent and identically distributed (IID) noise with variance σ_0^2, we have $\Sigma_0 = \sigma_0^2 I_n$, and it can be shown that, for a simple model with no coefficients for the Fourier series, when the change-point is located in the middle of a series of total length n, the variance of the estimator of the jump then writes:

$$var[\hat{\delta}_{GLS}] = \frac{4}{n^2}\sigma_0^2 \tag{4.11}$$

In the case when the noise is autocorrelated but remains homoscedastic, then $\Sigma_0 = \boldsymbol{VCV} = \boldsymbol{C}\sigma_0^2$, and $var[\hat{\delta}_{GLS}]$ is still proportional to σ_0^2. Finally, when the noise is autocorrelated and heteroscedastic, it can be shown that $var[\hat{\delta}_{GLS}]$ is proportional to the time-mean variance $< \sigma_t^2 >$. In the real data, this variance increases with the distance.

Figure 4.22 presents the graph of the estimated standard error of the jump estimator as a function of the time mean of the moving standard deviation for each series of difference. The relationship between the two variables is almost linear, as expected. The (positive) correlation between these two variables ranges from $r = 0.76$ to 0.84. For non-collocated series, the range of values is larger, where the larger values correspond to larger distances (distinguished by red and blue points). The collocated series, G-E and G'-E', have a maximum standard deviation of about 1.6 , whereas for non-collocated series, it lies between 3 and 3.5.

Figure 4.23 plots the noise level of the four non-collocated series as a function of the distance. As expected, the further away the nearby station is, more noise is introduced into the difference series. Consequently, a strong dependence of the uncertainty in the jump estimates on the distance becomes evident as depicted in Figure 4.24.

Figure 4.25 presents the t-values of the four non-collocatedseries plotted against distance.
Among these, G-G' and G-E' display a declining trend in t-values as distance increases. In contrast, G'-E and E-E' show a near uniform t-value distribution. This pattern is consistent with the findings observed in the t-values box plot presented in the paper. The decreasing trend in G-G' and G-E' is attributed to the rise in standard error with increased distance, while the uniform distribution in G'-E and E-E' is due to their relatively smaller jump amplitudes compared to the others.

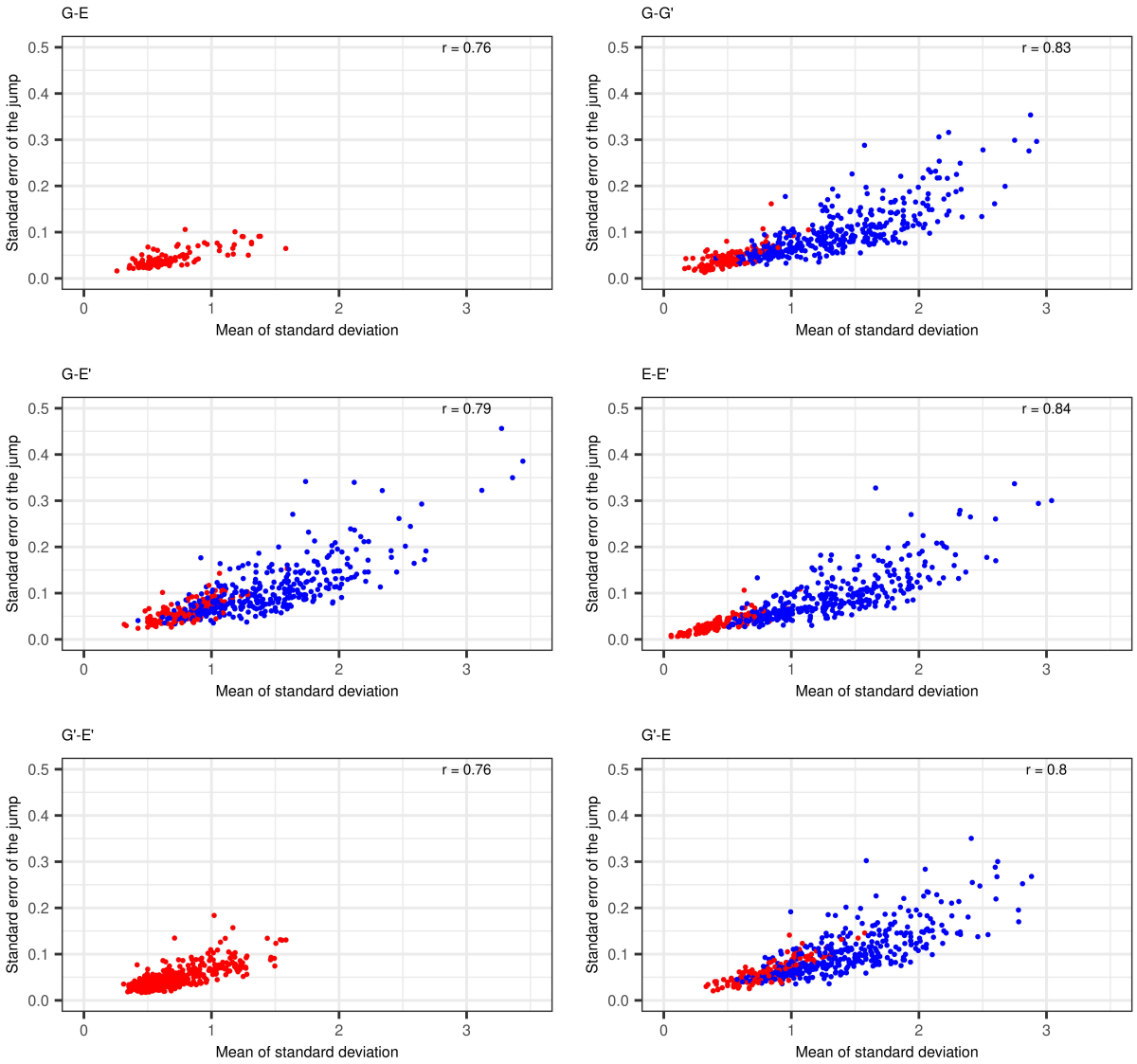


Figure 4.22 – Estimated standard error of the jump estimator as a function of the moving standard deviation for the four non-collocated series. The value r is the correlation coefficient between the two variables and the points are colored according to whether the distance between the series is less (red) or more (blue) than 50 km.

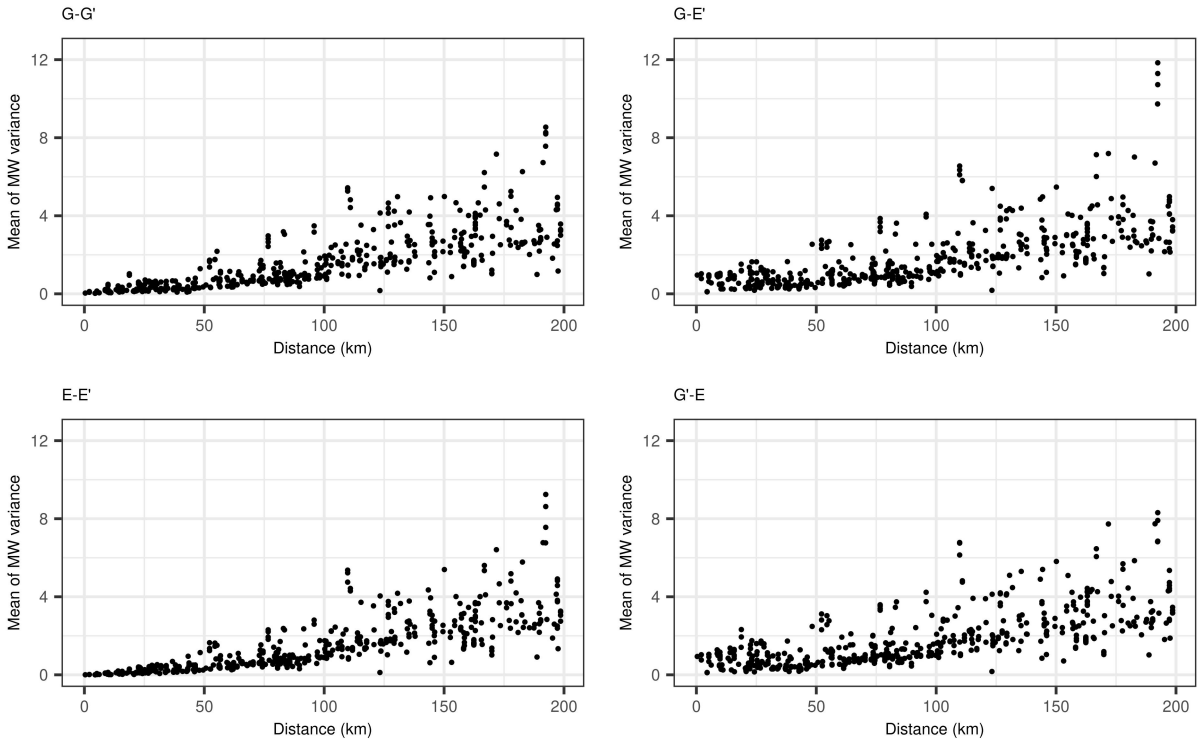


Figure 4.23 – Mean of the moving variance as a function of the distance in four non-collocated series.

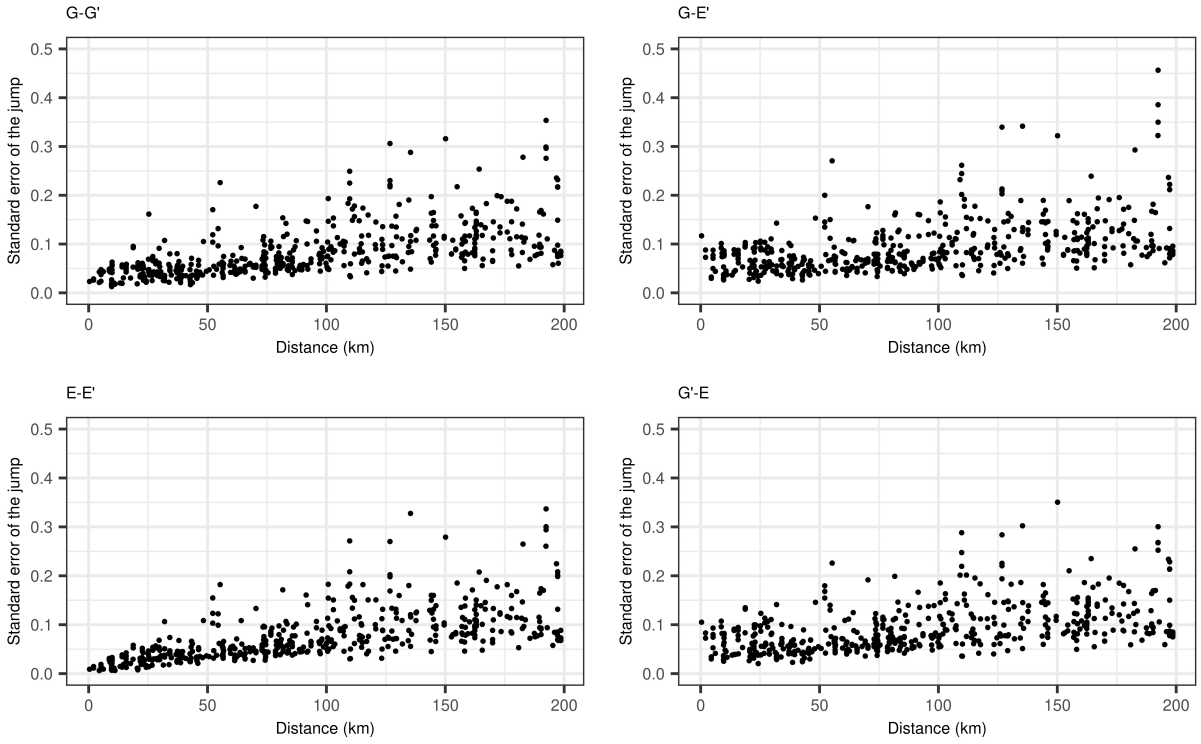


Figure 4.24 – Estimated standard error of the jump estimator as a function of the distance in four non-collocated series.

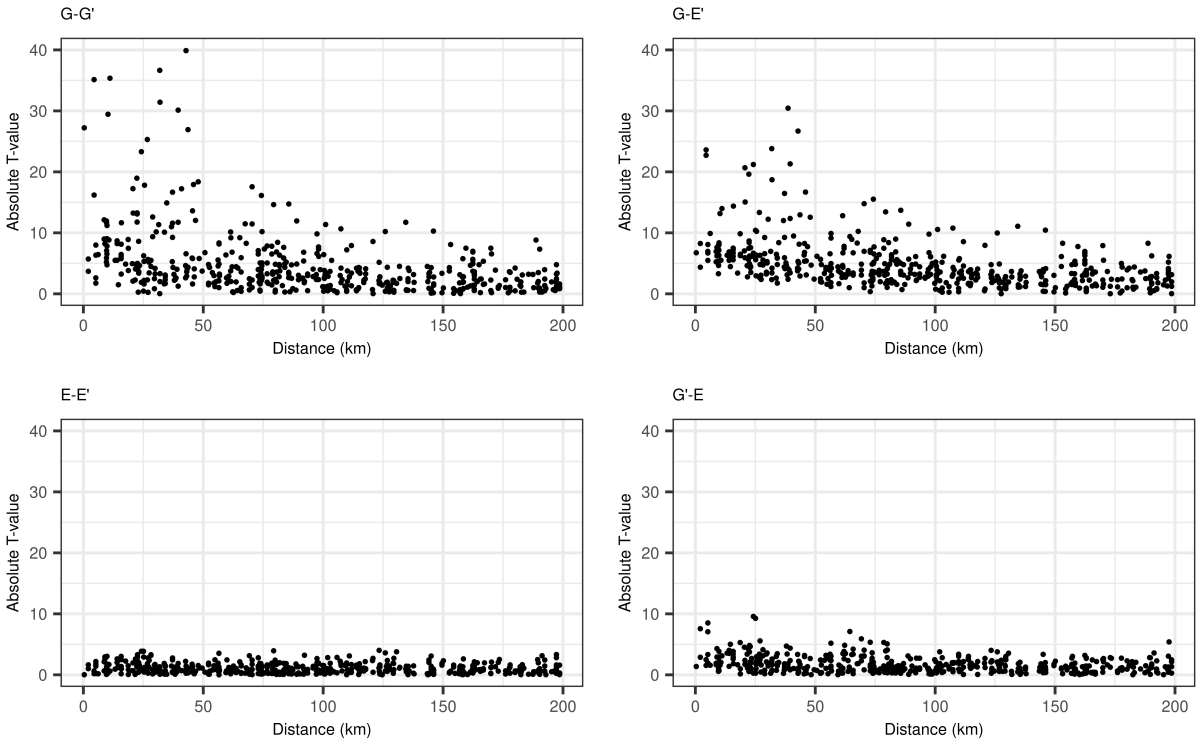


Figure 4.25 – T-value as a function of the distance in four non-collocated series.

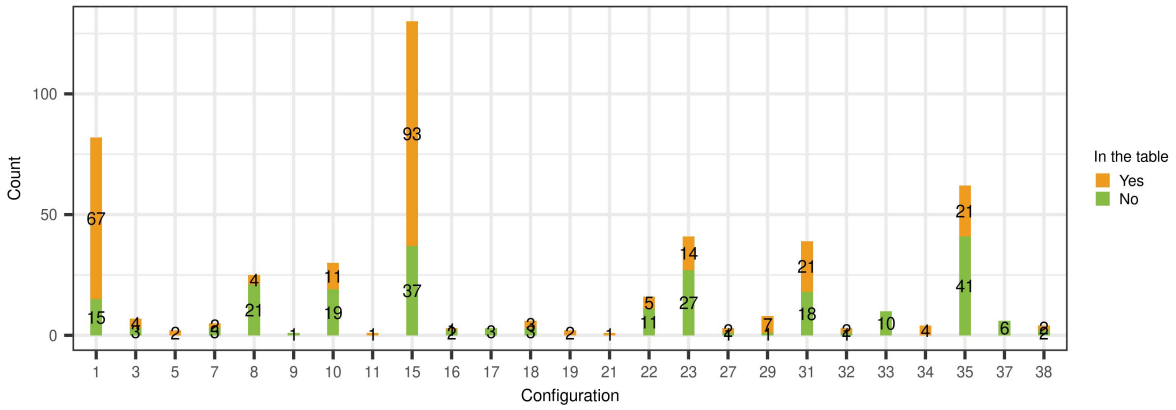
Predictive rule. Further insight into the attribution results presented in the subsection 4.1.5.3 of the paper can be gained by looking into the distribution of predicted configurations of the whole 494 cases before aggregation. Here we analyse three of the variants presented in the paper: (a) balanced and (c) imbalanced sampling with $\alpha = 0.05$, and (b) balanced sampling with $\alpha = 0.01$. The results are presented in Figure 4.26. For each configuration, we distinguish whether the test results from 6 series of differences corresponded to a configuration in the table or not.

At a significance level of $\alpha = 0.05$, 269 cases correspond to configurations in Table 4.1 (yellow bars in Figure 4.26a), predominantly in configurations c=15, 1, 31, 35, 23, and 10, which hold the highest conditional probabilities in the Test Table (Table 4.1). The green bars in the figure highlight cases that are not in the table and are predicted, for variant (a), as $c = 35$ (41 cases), $c = 15$ (37 cases), $c = 23$ (27 cases), $c = 8$ (21 cases), $c = 10$ (19 cases), $c = 31$ (18 cases), etc. The dominance of those configurations was previously explained through the three main groups in Table 4.5.

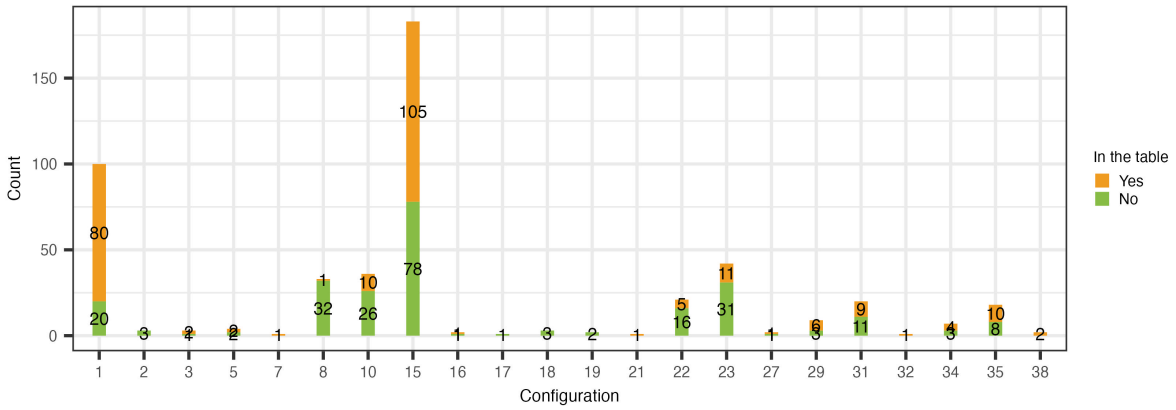
Modifying the significance level from 0.05 to 0.01 has a pronounced effect on the prediction results. We observe an increase in the cases predicted under configurations $c = 1$, 15, 8 and 22, while predictions for configurations $c = 31$ and $c = 35$ show a decline. This is consistent to changes observed in the final results in paragraph 4.1.5.3.

The final variant (c) is obtained by training with an imbalanced sample. To discern the influence of this imbalance consideration, we contrasted them with results from balanced sample. The results of both are given

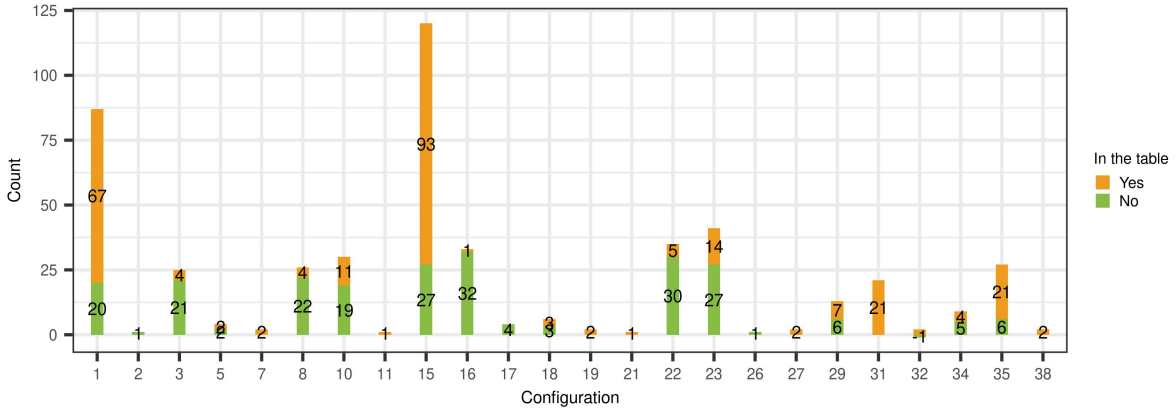
in Figure 4.26c and 4.26a respectively. While the count of the test results corresponding to a configuration, represented by orange bars, remains unchanged (since the significance level is consistent), the effects of sample balance manifest in three distinct points showed by green bars. Firstly, the G group ($c = 1, 15$) and E&E' group ($c = 10, 23$) maintain their values with only slight change. Secondly, configurations with simultaneous change-points at G and E possessing small marginal probabilities (e.g., configurations 31, 33 and 37) are absent in predictions, as expected by using imbalanced sample (see the discussion in the subsection 2.2.4.3). Thirdly, specific configurations, such as $c = 3$ and $c = 16$ (with test configurations (1,1,1,0,1,1) for 3 and (-1,-1,-1,-1,-1,0) for 16), show unexpectedly high prediction rates. These predictions arise from test combinations (-1,-1,-1,0,-1,0) and (1,1,1,0,1,0), which were predicted to configurations 35 and 31 in balanced sample training respectively (with test configurations (-1,-1,-1,0,-1,-1) for 35 and (1,1,1,0,1,1) for 31). The predictive rule of this variant modifies the E-E' from 0 to ± 1 diverging from the variant (a) that adjusted the G'-E from 0 to ± 1. This phenomenon arises due to the higher probability of the $c = 31, 35$ configurations.



(a) Training with balanced sample and $\alpha = 0.05$



(b) Training with balanced sample and $\alpha = 0.01$



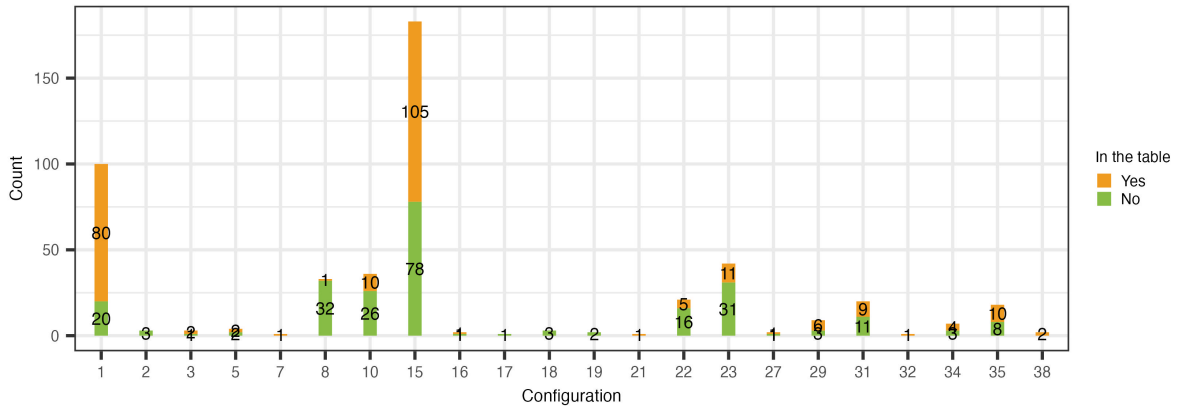
(c) Training with imbalanced sample and $\alpha = 0.05$

Figure 4.26 – Distribution of the predicted configuration for each of the 496 cases. Colors distinguish between test results aligning with table configurations (Yes in yellow) and those that do not (No in green).

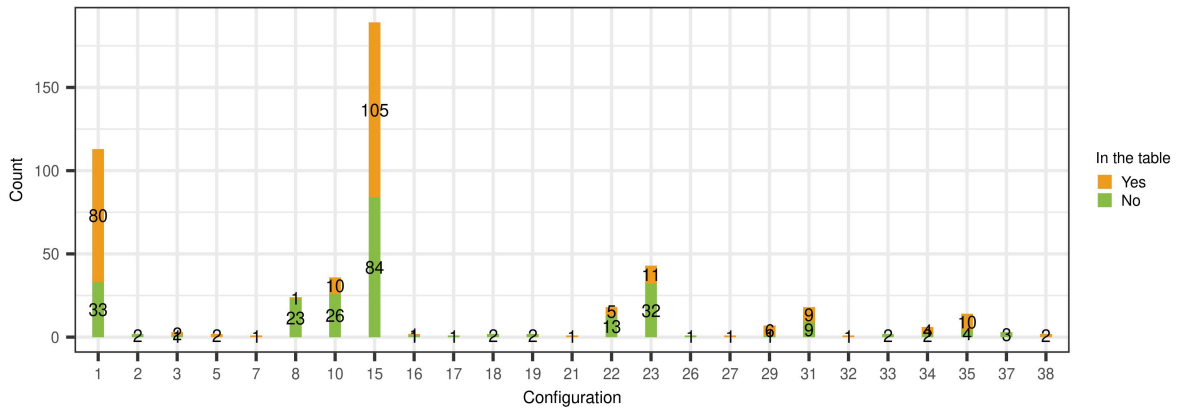
Discussion on the choice of the final classifier. The training process is applied $B = 20$ times (see subsection 4.1.5 of the paper) and the final classifier is chosen as the one with the best predictive power among the 20 classifiers. It is noteworthy that our dataset's limited size, with a lot of sparsity in terms of configurations, can lead to strong replication during the formation of the completed dataset, potentially causing overfitting issues

in the construction of the classifier. Consequently, 2 different classifiers among the 20 can both show similar good performance in term of mean misclassification error, but they may predict exactly the same results. The question here is to assess to which extent choosing one classifier instead of another impacts the prediction results.

Figure 4.27 illustrates the outcomes from two distinct Random Forest classifiers (for the repetitions $b = 3$ and $b = 4$) on the real dataset, both achieving an identical mean error of 0. These classifiers were trained with an equal sample size of $R = 100$ for each configuration, at a significance level of $\alpha = 0.01$. We observed some difference: configurations $1, 8, 15$, and 23 exhibit varying numbers of cases, and some other configurations, such as $26, 33$, and 37, appear or disappear after the classification obtained in $b = 4$. This highlights how sensitive the results are to the sampling process, especially in our case where dataset is of small size.



(a) Repetition b=3



(b) Repetition b=4

Figure 4.27 – Distribution of the predicted configuration for each of the 496 cases for two predictors constructed during two repetitions b of the algorithm in the case of an equal training sample size of $R = 100$ and with a significance level of $\alpha = 0.01$ for the test results. Colors distinguish between test results aligning with the Test Table (Table 4.1) (Yes in yellow) and those that do not (No in green).

4.2.3 Strategies for further improving the prediction results

4.2.3.1 Correction of distance bias

As highlighted in the previous subsection, the distance strongly influences the noise level in difference series and consequently the jump uncertainty. We propose a method to correct biases in the t-values of the non-collocated E-E' series. According to the rule n°2 presented in the paper, the series E and E' should be the same due to the large spatial correlation in the reanalysis. Any jump observed in E-E' is a manifestation of climatic differences between the locations of the two stations which may induce false detections in the non-collocated series. The idea is to estimate this bias from the E'E' series and correct the other series for it. This correction can potentially change the outcome of the significance test. If effective, we would expect an increase in significant jumps in G-E' and a decrease in G'-E, assuming that the change-point actually comes from G. Figure 4.28 shows the evolution of significance levels in three tests: after correction, G-E' sees an increase in significant tests, while on the contrary G'-E has more insignificant tests. These are slight changes but coincide with our expectations.

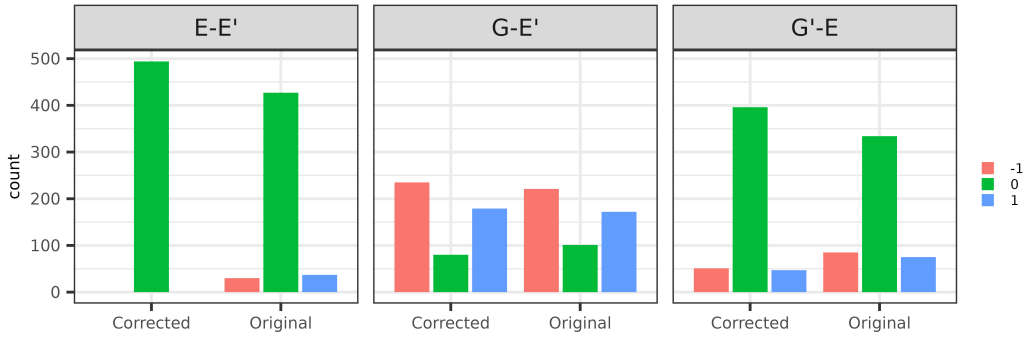


Figure 4.28 – Test count in three groups: 0 indicates no jump, 1 signifies a significant upward jump, and -1 represents a significant downward jump, from original results and after bias correction.

Using this strategy reduces drastically the number of initial possible configurations from 36 to 10 because the E-E' test is now removed. The remaining configurations are: 1, 5, 10, 13, 15, 18, 23, 27, 31 and 35 (configurations containing a 0 for E-E'). This modification has a small positive effect on the test results. Figure 4.29 presents the configurations from the test results before and after this correction. The number of cases where the test results correspond to a configuration in the table increases despite the substantial reduction in rows. Number of cases initially match to configuration 1 and 15 increase. This shift can be attributed to the previously mentioned increase in significant tests in G-E' and a decrease in significant tests in G'-E.

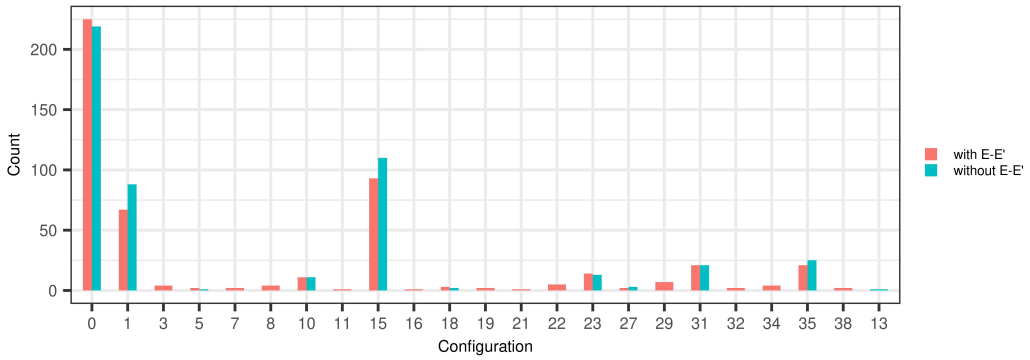


Figure 4.29 – Configuration count before prediction: when considering the series E-E' and when ignoring it. A value of 0 signifies test results not present in the table.

4.2.3.2 Multiple testing correction

The attribution method is based on the on simultaneous testing results from six series, which poses a multiple testing problem. This well-known issue posits that an increase in the number of tests made leads to a higher probability of erroneous conclusions. Simply, under the null hypothesis, the probability of a false rejection is denoted as α and thus, the probability of a true detection is $1-\alpha$. For k independent tests, the rate of correctly retaining the null hypothesis is $(1-\alpha)^k$. Consequently, the probability of rejecting at least one test when all null hypotheses are true is expressed as $1-(1-\alpha)^k$, which called family wise error rate (FWER). The commonly used Bonferroni correction controls the FWER. If we test each hypothesis at a significance level of α, we guarantee that the probability of having one or more false positives is less than α. However, this control is often too strict. An other less restrictive measure is the False Discovery Rate (FDR) proposed by Benjamini and Hochberg (1995). It is defined as the proportion of false discoveries:

$$FDR = E[V/R], \tag{4.12}$$

where R represents the total number of rejected null hypotheses or positive and V represents the number of erroneously rejected null hypotheses or false positive. When R is equal to V, FDR becomes FWER. Control for FDR at level α consists in: (1) calculating the p-values obtained from the k tests ordered from smallest to largest ($p_{(1)} \leq p_{(2)} .. \leq p_{(k)}$), (2) finding the test with the highest rank j for which $p_{(j)} \leq \frac{j}{k}\alpha$ and (3) declaring the tests of rank $1, 2, \ldots, j$ as significant.

Figure 4.30 compares the significance of the tests for each series of difference before and after the FDR correction. As expected, we observe an increase, but a slight one, of the insignificant test, except obviously for G-E. The effect is minor, but it can have a positive impact on the configuration obtained from the test results. These configuration numbers are given in Figure 4.31. Firstly, after the correction, there is a slight increase in the number of cases falling into configurations 1 and 15, while configurations 8, 22, and 31 show a slight decrease. An advantageous feature of FDR is its consideration of G'-E', which exhibits more significant jumps due to lower noise in comparison to other non-collocated series (depicted in Figure 4.22). Secondly, subsequent to the correction, there is a rise in the count of cases where the six tests align with a configuration in the table.

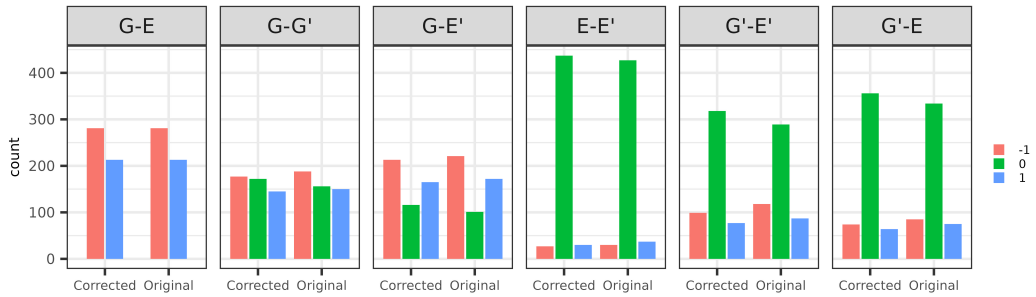


Figure 4.30 – Test count in three groups: 0 indicates no jump, 1 signifies a significant upward jump, and -1 represents a significant downward jump, from original results and after FDR correction.

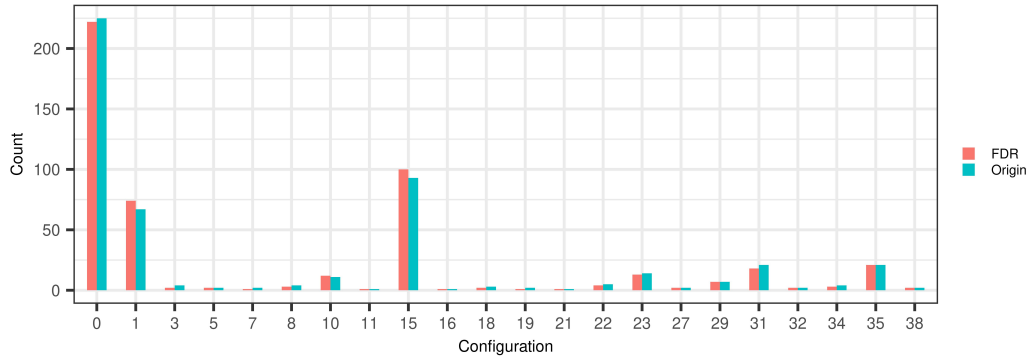


Figure 4.31 – Configuration count before prediction from original results and after FDR correction. A value of 0 signifies test results not present in the table.

4.2.4 Conclusions

In this section, we have carried out additional analyzes in order to detail and discuss in more depth some of the results of the article (in particular on the identification of models and on the different stages of attribution, i.e. on the tests and the predictive rule).

Firstly, we detailed the results for studying the performance of the method used to characterize the data through simulations. The results revealed a relationship between the accurate identification of the noise model and the length of the series or the parameter values. Specifically, the ARMA(1,1) noise model requires a longer series for accurate identification compared to AR(1) and MA(1) models. Additionally, when the parameter is small (<0.3), there is a higher likelihood of misidentifying the AR(1) model as MA(1) or vice versa.Furthermore, the presence of data gaps can intensify model identification confusion and parameter estimate dispersion, but it has been shown that the impact remains small for reasonable fraction of gaps (e.g. < 20%, see 4.19).

Secondly, we analyzed in more details the prediction results from the significance tests, categorizing them based on their presence in the Test Table (Table 4.1 of the paper). Notably, the standard error of the jump estimates, and thus the test result, strongly depends on the data noise, which is influenced by the distance between the two series involved in the test. This relationship is linear for the G-E, G-G', and G-E' but not for

the E-E', G'-E, and G'-E' series. This is simply because jumps are more likely to be due to G rather than E, G', and E', as previously mentioned in the paper result.

Finally, we tested potential improvements to the prediction step by making bias correction and applying multiple testing correction. Both proposed modification yield positive, although, marginal improvements, such as a slight increase in the number of high probability configurations and a slight decrease in some low probability configurations. Because the improvements are only marginal, they are not further considered.

Chapter 5

Conclusions and perspectives

5.1 Conclusions

Water vapor plays a pivotal role in climate change due to its powerful feedback mechanism. Understanding and quantifying the change in water vapor within the context of climate change represent a prominent scientific challenge, as emphasized in recent research (Colman and Soden, 2021; Douville et al., 2022). While there is a global consensus regarding an upward trend in water vapor, several studies have noted discrepancies in trend estimates, particularly at the regional level, when comparing data from climate models, reanalyses, and observational sources (Parracho et al., 2018; Santer et al., 2021; Allan et al., 2022).

In this context, GNSS provides reliable long-term IWV data that are available under all weather conditions and independent from the reanalyses, which are useful to assess and validate the accuracy of reanalyses and climate models, thereby facilitating enhancements in these models. However, the presence of abrupt offsets in the GNSS data is problematic since trend estimates are sensitive to such inhomogeneities.

Within the overarching goal of providing a homogenized long term record of IWV data from GNSS, this study utilized a relative homogenization approach. The comprehensive homogenization process comprises four fundamental steps:

- firstly, generating the differences in IWV between GNSS and reanalysis data (used as a reference);
- secondly, segmenting this series;
- thirdly, attributing the identified change-points to either GNSS or reanalysis series;
- finally, correcting the IWV GNSS time series.

For the second step of detecting change-point, we applied the GNSSseg segmentation method developed by Quarello (2020) specifically for these data.

This thesis made contributions in two key areas. First, it provided enhanced insight on the sensitivity of the GNSSseg segmentation method and trend estimates to the GNSS and reference (reanalysis) data properties. Second, an innovative automatic attribution method that was developed, which is a crucial step, with the segmentation, in the global homogenization process.

Sensitivity of the segmentation method and trend estimates to data properties. This study was presented in the first paper (Chapter 3). The comparison of segmentation results of 81 common stations in two GNSS datasets (IGS repro1 and CODE REPRO2015) and two reanalyses (ERA Interim and ERA5) revealed the sensitivity of the segmentation method to different data properties. In particular, the segmentation method was found to be sensitive to variations in noise magnitude (heteroscedasticity) and the presence of periodic bias in the data. When we exchanged either the GNSS IWV dataset or the reference dataset, it was observed that only about half of the detected change-points remained similar. These sensitivities are mainly explained by the differences in the heteroscedasticity and biases in the GNSS and reanalysis data. Especially, it was shown that more recent GNSS processing or reanalysis data exhibit reduced noise and biases, and thus smaller representativeness differences between them. This consequently enabled the segmentation method to identify smaller jumps in the IWV difference series. Modification in the series length or auxiliary data, which includes surface pressure and the weighted mean temperature used in the conversion from ZTD to IWV, resulted in a much more substantial similarity of $70 - 80\%$ in change-point detection. These factors subsequently influenced trend estimates, as evidenced by the mean trend and its dispersion.

We also observed a substantial impact of jump corrections on estimated trends. Notably, the uncertainty in the estimated trends resulting from the use of different reprocessed GNSS datasets or reference reanalyses in the segmentation ranged from 0.015-0.019 kg m^{-2} $year^{-1}$ (RMS difference between the tested pairs of data sets) when only the validated change-points were used and 0.002 to 0.012 kg m^{-2} $year^{-1}$ when all the change-points were used. However, on a global scale, homogenized GNSS data (corrected using either validated or all change-points) and both reanalyses agreed on a global mean IWV trend estimate of 0.027-0.030 kg m^{-2} $year^{-1}$, indicating a global moistening trend close to 7% K^{-1} over the past 2.5 decades. The longer time series (25 years) provided higher accuracy in trend estimates, with a mean standard error of 0.018 kg m^{-2} $year^{-1}$ and a dispersion of approximately 0.03 kg m^{-2} $year^{-1}$ across the global network.

This study also addresses two main limitations in the homogenization process. First, regarding the segmentation method, we observed that some stations exhibit interannual and longer-term variations in the noise and bias characteristics which are not currently modeled in the segmentation method (both the noise variance and the bias are assumed periodic with a fundamental period of 1 year). At those sites, interannual and/or decadal variations in bias especially were captured by the segmentation. The long-term variations could be reduced, for example, through the use of a better reference reanalysis or a nearby GNSS station, rather than a reanalysis. Second, only about 35% of change-points are validated by metadata, meaning that 65% of the detections can be undocumented changes in GNSS or changes in reanalysis data. This highlights the critical role of the attribution step, which aims at determining the origins of identified change-points.

Development of an automatic attribution method. The method developed in this thesis relies on a the combination of GNSS and reanalysis data from a main station with similar data from nearby stations. For each detected jump in the main series, and each associated nearby station, six series of differences are formed from the four base series and a test of significance is performed on the jump. The development of the proposed method involves several meticulous steps. Firstly, we characterize the heteroscedasticity and autocorrelation structure in the six series of differences. The considered noise models are white noise, AR(1), MA(1), and ARMA(1,1). Notably, we show that only a tiny number of series are identified as white noise. Secondly,

we evaluate the significance of jumps in each individual series using the generalized least squares (GLS) approach, with the identified noise model, which provides statistically more efficient error estimators. Finally, we construct a predictive rule to make the attribution decision. The novelties of our approach lie in the use of the GLS estimation in this context, and in the construction of a predictive rule, using test results obtained from real data with a resampling procedure taking into account the small samples of our dataset (here limited to 494 cases). The method is described in detail in Chapter 4. When applied to the set of 81 GNSS stations from the CODE REPRO2015 dataset, with ERA5 reanalysis used as reference and the NGL GNSS dataset for the nearby stations, 62% of the tested change-points were attributed to GNSS, 19% to the reanalysis, and 10% to changes in both GNSS and reanalysis. Detailed analysis of the GLS test results input to the prediction rule helped us to understand the sensitivity of the attribution method and interpret the final result after aggregation (i.e. when predicted results from several nearby stations are combined). Indeed, the prediction results are sensitive to the resampling strategy used for the construction of the predictive rule. We tested several variants, using different significance levels for the tests, balanced or imbalanced samples and different aggregation rules (based on case-specific prior probabilities from the Test Table, Table 4.1, or using the distance to the nearby station). The imbalance of configurations observed or assumed in reality seems better taken into account with a balanced resampling strategy combined with a aggregated rule based on the prior probabilities. Test results are also sensitive to the magnitude of noise in the data which depends on the distance between the main and nearby stations. The noise levels increase substantially with the distance between the main and nearby stations in a sparse network. This high noise can hinder the significance tests in the four non-collocated series, thus impacting the final prediction result. This challenge can be addressed by applying our method within denser networks, such as provided in NGL dataset. The potential of this dataset will be explored in the near future. Some preliminary results are already presented in the Appendix.

5.2 Perspectives

This study raises certain questions that could serve as potential avenues for future work aimed at enhancing GNSS data processing and climatic homogenization.

a) In the context of GNSS data processing, segmentation proves to be highly sensitive to small changes in mean of the series. These characteristics are influenced by several key aspects of GNSS data processing, with particular emphasis on factors such as: the a priori ZHD correction, the antena/radome calibration model, the mapping function and the elevation cutoff angle. The consistency and reliability of GNSS products generated through different software packages and incorporating varying processing features can be a subject of scrutiny. This can be addressed by conducting comparative analyses of different GNSS tropospheric products, especially focusing on noise, bias and homogeneity. The assessment of data homogeneity can be accomplished through segmentation (to detect changes in the mean) and attribution methods (to identify unambiguously those of GNSS origin). Afterwards, it may be tempted to optimize the GNSS processing in order to make the products intrinsically more homogeneous. The homogenization tool can thus serve to assess the progress accomplished at the processing level, in addition to completing the final homogenization of the reprocessed data.

b) In the following, we outline ideas to enhance our homogenization method, with a particular emphasis on

improving both the segmentation and the attribution methods.

The first goal in the short term is to test the current attribution method with an extended GNSS dataset from the NGL repro3. With this dataset, we can strengthen the nearby selection rules by selecting nearby series with a smaller percentage of gaps and shortening the distance between the main station and nearby stations. This will reduce the magnitude of the noise in the difference series and improve the power of the test procedure.
Furthermore, as mentioned in section 4.1.5, the synthetic dataset is generated from the test results of real data. With the NGL repro3 dataset, in particular, the larger size of the dataset can improve the prediction rule. Firstly, it can improve the performance of the current decision rule proposed in chapter 4 because with more tested changes there will be fewer repetitions in the synthetic dataset constructed by resampling. Secondly, we could consider a new method to construct this decision rule. For example, rather than doing resampling series by series, we could consider doing resampling test by test (on all the results at the same time), which would be fairer to take into account the dependence between the tests.

However, one critical issue arises with the NGL data regarding the validation of the final results with metadata. The preliminary results presented in A show a drop in the percentage of validated change-points which is explained by the fact that only one half of the analysed stations have metadata reported in the NGL data holding. This could be partly improved by running additional quality control procedures based on the RINEX files as was done by Quarello et al. (2022).

Second, a R package of the attribution method will be made available, enhancing its accessibility to the scientific community.

Other approaches to the segmentation method implemented currently could be tested, such as using nearby GNSS stations as references instead of a reanalysis. When both the main and nearby stations are processed in a consistent manner, the remaining inhomogeneities would mainly originate from station-specific changes, such as instrumental changes or changes in the surrounding environment. Consequently, this approach can facilitate the attribution of change-points and lead to significant improvements in the homogenized series.

The results of heteroscedasticity and autocorrelation in the difference series, as demonstrated in Chapter 4, underscores the need for two primary enhancements to the segmentation method. Firstly, the current practice of modeling heteroscedasticity on a monthly basis within the segmentation method may benefit from being replaced by a more comprehensive continuous function, as exemplified in Figure 4.2. Secondly, the current use of white noise does not take into account the underlying autocorrelation patterns present in the data. A more appropriate model for data autocorrelation (e.g., AR(1), MA(1), or ARMA(1,1)) should be used to improve the segmentation method. Finally, using an appropriate loss function for segmentation inference like Hubert or Biweight loss (instead of log-likelihood) could avoid data pre-processing with respect to outliers.

In the long-term exploration of this topic, conducting an analysis of homogenized GNSS IWV trends presents a compelling opportunity, especially within the context of climate change studies. This analysis provides empirical evidence to address a key question: How will water vapor change in a warmer climate? The

water vapor feedback is frequently estimated using a global climate model (GCM), which strongly depends on parameterization. Homogenized GNSS datasets can be compared and used as a reference for tuning parameters in climate models, aiding in obtaining the "best" estimates of water vapor feedback and its uncertainties.

Appendix A

Appendix: Preliminary results with an extended dataset

This appendix presents some preliminary results of the application of the full homogenization process to an extended GNSS dataset of more than 6000 stations extracted from the NGL data holding. At the time of writing, several steps are already completed, including station selection, ZTD data conversion and screening, segmentation, and validation with respect to the NGL metadata. The attribution step is currently pending. After the attribution, the correction of jumps in the GNSS series will be carried out and trends will be estimated. The IWV data before homogenization is already available publicly here `https://doi.org/10.25326/518`. The homogenized version will be added later.

⋆ Station selection

The NGL repro3 data holding contained 20,747 stations, as of 13 April, 2023. It is updated on a daily basis. Data products are available in the form of time series of daily station coordinates and daily SINEX tropospheric (tropo) files containing the estimated ZTD and gradients parameters with 5-min sampling. Station metadata is provided in the form of a station list, including a 4-character station ID (renamed from their original names when they are already registred from a previous data network/provider), a priori station coordinates, dates of start and end, and number of computed days.

For the purpose of climate trend estimation, a sub-set of stations with more than 10 years of observations and less than 10% of gaps (at the daily sampling rate) was selected. This selection came up with 6197 stations. After checking the station position stability from the coordinate time series, with thresholds 0.001° in latitude and longitude, and 1 m in height, the number is decreased to 6058 stations. The 5-min resolution tropo files have been downloaded for this subset of stations, and ZTD estimates and their formal errors have been extracted are averaged into daily values.

⋆ Data screening

Daily ZTD values were screened for outliers using a new procedure compared to the earlier procedure used with IGS repro1 and CODE REPRO2015. Firstly, the formal errors have been passed through a segmentation algorithm and a post-processing was developed to detect both singular outliers (e.g. one day with a high formal error)

and extended periods (weeks to months) of high formal errors which are symptomatic of degradation of GNSS instrumentation or measurement conditions. In some occasions, long-term drifts (over several years) of formal errors were also discovered and four stations were blacklisted for this reason. This post-processing relies on several conditional tests and thresholds which required some tuning before being applied to the data set. 0.62% of the ZTD values and 6 stations were rejected based on this screening procedure based on the formal errors only.

Secondly, a complementary screening step was applied based on the GNSS - ERA5 IWV difference. The ZTD values were beforehand converted to IWV using the same procedure as for the IGS repro1 and CODE REPRO2015 data sets, except that here ZHD and T_m values computed from the model were first averaged into daily values and applied to the daily GNSS ZTD estimates. The screening method used for the IWV differences also differs from previous procedures in the sense that it is applied on the full time series (instead of yearly segments) which allows to implement a sliding window approach where the individual daily values are compared to a moving median and tested with respect to a robust scale estimator (MAD). This procedure also relies on some parameterization (e.g. window size and test thresholds) that had to be tuned. With the current settings, it rejects an additional 0.89% of daily values and 2 additional stations, leaving 6048 stations available for the application.

- ⋆ Segmentation

The quality-checked IWV differences were passed through the GNSSseg segmentation method with standard settings (as in Quarello (2020) and Nguyen et al. (2021)). The segmentation detected 15680 change-points in 6048 stations, which represents an average of 2.6 change-points per station. Among these, 611 stations (10.1%) were considered homogeneous (no change-point detected), and 5437 stations (89.9%) had at least one change-point. The maximum number of detected change-points in one series was 16 (station J137).

The post-processing analysis described in Quarello (2020) was applied to detect "clusters" (groups of consecutive segments with significantly higher means than their adjacent segments) using the standard settings (distance between change-points of 80 days and significance level of 0.05). It detected 1427 clusters among which 1005 deemed significant. For each significant cluster, all the change-points pertaining this cluster were replaced by one single change-point. For the non-significant clusters, all the change-points were eliminated. After this step, 13923 change-points were remaining.

- ⋆ Validation with respect to metadata

The metadata provided by NGL (named "steps" file) contains information on changes in station instrumentation (antenna, receiver, radome) and other notable information (site change, monument change, volcanic eruption) as well as events flagged as "unknown". In total, we regrouped the different types of changes into 11 categories. However, we must emphasize that information is registered by NGL only for 6283 stations out of 20747. Among the 6048 stations that were analysed, only 3309 (49.6%) have metadata reported. The most frequent changes involve antennas, receivers, cutoff changes, and radomes, in decreasing order, which altogether count for 98% of the registered changes.

Among the 13923 detected change-points, 6921 change-points could be confronted to metadata and 1228 (8.8%) were coincident or "validated" with respect to known changes within +/- 62 days (this is the same window as used in Quarello et al. (2022) and Nguyen et al. (2021)). The proportion of validated antenna, receiver, cutoff, and radome changes is 71.7%, 35.7%, 20.4%, and 34.7%, respectively (some changes involve more than one type, so the sum exceeds 100%), which represent again 98% of all validated changes. We note that the percentage of validated changes (8.8%) of all detected change-points is significantly lower than what was found with the IGS repro1 and CODE REPRO2015 data reported in paper No. 1. The reason is that the metadata provided by NGL are available for only one half (49.6%) of the analysed stations. This points to a severe lack of essential information necessary to control the quality and validate the segmentation results of this dataset. Hopefully, the attribution method will be able to achieve more robust conclusions on the significance and origin of the detected change-points.

⋆ Attribution

At the time of writing, the attribution of the NGL dataset was not yet completed. Several preliminary steps are currently being performed. The first one concerns the selection of the main and nearby stations. Because the ultimate goal is to estimate climatic trends, we decided to select as main stations all those having at least 20 years of IWV data. This selection retains 718 candidate main stations, among which 20 are homogeneous and 698 have at least one change-point.

Then we select for each main station all possible nearby stations within a maximum horizontal distance of 200, 100, or 50 km, and a vertical distance smaller than 500 m. With these distance limits, some of the main stations have no nearby station and their change-points cannot be tested with the attribution method. The number of main stations having at least one nearby stations decreases from 698 to 673, 631, and 567, respectively, for the three maximum distances. Next we checked more precisely how many of the detected change-points in the main stations can actually be tested based on the matching of the time periods with the nearby stations. Therefore, we pre-select the change-points, for each main station, which do not have another change-point closer than 250 "effective" days (by "effective" we mean that only days with data are counted here, i.e. days with missing data are not counted), either within the time series of the main station itself or in the considered nearby station. This step rejects some of the main change-points and rejects also some of the candidate nearby stations when they have (nearly) coincident change-points. This pre-selection gives the following results (this is only for main stations which have at least one change-point):

- With the 200 km distance limit: 632 main stations and 1708 change-points can be tested with the help of 3960 nearby stations (total 44474 main-break-nearby triplets). Among them, 407 main stations can be tested for all their change-points (i.e., can be fully homogenized). The number of nearby stations per change-points is: mean = 26.0, min = 1, max = 152.
- With the 100 km distance limit: 581 main stations and 1455 change-points can be tested with the help of 2467 nearby stations (total 17208 main-break-nearby triplets). Among them, 317 main stations can be tested for all their change-points (i.e., can be fully homogenized). The number of nearby stations per change-points is: mean = 11.8, min = 1, max = 79.
- With the 50 km distance limit: 499 main stations and 1115 change-points can be tested with the help of 1302 nearby stations (total 6722 main-break-nearby triplets). Among them, 202 main stations can be

tested for all their change-points (i.e., can be fully homogenized). The number of nearby stations per change-points is: mean = 6.0, min = 1, max = 50.

These statistics indicate that on average between 6 and 26 nearby stations can be used to test change-points in the main stations, depending on the distance limit, and that a good fraction of the main stations (between 40 and 64%) can be fully homogenized. For those change-points that cannot be tested, a manual validation remains necessary, which can be assisted by the use of metadata when available.

The issue of lacking metadata in the NGL data holding can be partly improved by running additional quality control procedures based on the RINEX files as was done by Quarello et al. (2022). Fortunately, the NGL data holding contains Quality Assurance (QA) files which contain statistics computed from the GipsyX processing outputs, for each station, daily. The analysis of these statistics may be of some utility to complement the existing metadata for the validation. For example, if a change-point detected by the segmentation of the IWV difference series coincides with a shift in one of the data quality metrics provided in the QA files, this may give some credit to consider the change-point are originating the GNSS series. Note that the segmentation of formal errors implemented in the new screening method described above also follows this idea and these results may be used as well, as are change-points detected in the GNSS position time series.

The next step before the attribution is the vertical correction of all main-nearby couples. This accuracy of this correction is also in the process of being improved compared to the version that was used in paper No. 2. The procedure will be the similar and apply a 2-parameter correction model as proposed by Bock 2022, but the coefficients will be computed from ERA5 reanalysis on the native 0.25° grid from the four surrounding columns for each station (as opposed to the use of a global 2.5° grid in paper No. 2). Another improvement will be the use of a sliding window regression with a 1-day increment (i.e. the coefficient will be regressed from ERA5 data in a window for each central day) instead of a monthly regression time step.

Bibliography

Alexandersson, H. (1986). "A homogeneity test applied to precipitation data". In: *Journal of Climatology* 6.6, pp. 661–675. DOI: 10.1002/joc.3370060607. URL: https://doi.org/10.1002/joc.3370060607.

Allan, R. P., K. M. Willett, V. O. John, and T. Trent (2022). "Global Changes in Water Vapor 1979–2020". In: *Journal of Geophysical Research: Atmospheres* 127.12. e2022JD036728 2022JD036728, e2022JD036728. DOI: https://doi.org/10.1029/2022JD036728. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022JD036728. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022JD036728.

Andrews, D. W. K. (1991). "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation". In: *Econometrica* 59, pp. 817–858.

Arlot, S. (2018). *Validation croisée*.

Bengtsson, L. (Apr. 2010). "The global atmospheric water cycle". In: *Environmental Research Letters* 5.2, p. 025202. DOI: 10.1088/1748-9326/5/2/025202. URL: https://doi.org/10.1088/1748-9326/5/2/025202.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the royal statistical society series b-methodological* 57, pp. 289–300.

Bevis, M., S. Businger, S. Chiswell, T. Herring, R. Anthes, C. Rocken, and R. Ware (1994). "GPS Meteorology: Mapping Zenith Wet Delays onto Precipitable Water". In: *J. Appl. Meteorol.* 33, pp. 379–386.

Bevis, M., S. Bussinger, T. A. Herring, C. Rocken, R. A. Anthes, and R. H. Ware (1992). "GPS Meteorology: Remote Sensing of Atmospheric Water Vapor Using the Global Positioning System". In: *J. Geophys. Res.* 97, pp. 15787–15801.

Birgé, L. and P. Massart (2001). "Gaussian model selection". In: *J. Eur. Math. Soc.*, pp. 203–268.

Blewitt, G., W. C. Hammond, and C. Kreemer (Sept. 2018). "Harnessing the GPS Data Explosion for Interdisciplinary Science". In: *Eos* 99. DOI: 10.1029/2018eo104623. URL: https://doi.org/10.1029/2018eo104623.

Bock, O. (2016). *GPS data: Daily and monthly reprocessed IWV data from 120 global GPS stations, version 1.2*. DOI: 10.14768/06337394-73a9-407c-9997-0e380dac5591.

— (2019). *Global GNSS IWV data at 436 stations over the 1994-2018 period*. DOI: 10.25326/18.

— (2020a). "Standardization of ZTD screening and IWV conversion". In: *Advanced GNSS Tropospheric Products for Monitoring Severe Weather Events and Climate: COST Action ES1206 Final Action Dissemination Report*. Ed. by J. Jones, G. Guerova, J. Douša, G. Dick, S. de Haan, E. Pottiaux, O. Bock, R. Pacione, and R. van Malderen. Springer International Publishing. Chap. 5, pp. 314–324. ISBN: 9783030139018. DOI: 10.1007/978-3-030-13901-8_5.

Bock, O. and A. Parracho (2019). “Consistency and representativeness of integrated water vapour from ground-based GPS observations and ERA-Interim reanalysis”. In: *Atmos. Chem. Phys.* 19, 9453–9468. DOI: 10.5194/acp-19-9453-2019.

Bock, O. (July 2012). “GNSS: géodésie, météorologie et climat”. version corrigée du 15/08/2013. Habilitation à diriger des recherches. Université Pierre et Marie Curie - Paris VI. URL: https://theses.hal.science/tel-00851617.

— (2020b). *Global GNSS Integrated Water Vapour data, 1994-2020*. en. DOI: 10.25326/68. URL: https://en.aeris-data.fr/landing-page/?uuid=df7cf172-31fb-4d17-8f00-1a9293eb3b95.

Bock, O., P. Bosser, and C. Mears (Oct. 2022). “An improved vertical correction method for the inter-comparison and inter-validation of integrated water vapour measurements”. In: *Atmospheric Measurement Techniques* 15.19, pp. 5643–5665. DOI: 10.5194/amt-15-5643-2022. URL: https://doi.org/10.5194/amt-15-5643-2022.

Bock, O., X. Collilieux, F. Guillamon, E. Lebarbier, and C. Pascal (May 2019). “A breakpoint detection in the mean model with heterogeneous variance on fixed time intervals”. In: *Statistics and Computing* 30.1, pp. 195–207. DOI: 10.1007/s11222-019-09853-5. URL: https://doi.org/10.1007/s11222-019-09853-5.

Bock, O., P. Willis, M. Lacarra, and P. Bosser (Dec. 2010). “An inter-comparison of zenith tropospheric delays derived from DORIS and GPS data”. In: *Advances in Space Research* 46.12, pp. 1648–1660. DOI: 10.1016/j.asr.2010.05.018. URL: https://doi.org/10.1016/j.asr.2010.05.018.

Bock, O., P. Willis, J. Wang, and C. Mears (2014). “A high-quality, homogenized, global, long-term (1993–2008) DORIS precipitable water data set for climate monitoring and model verification”. In: *J. Geophys. Res. : Atmos.* 119.12, pp. 7209–7230. DOI: 10.1002/2013JD021124.

Bock O. Pacione R. Ahmed F. Araszkiewicz A. BaŁdysz, Z. et al. (2020). “Use of GNSS Tropospheric Products for Climate Monitoring (Working Group 3)”. In: *Advanced GNSS Tropospheric Products for Monitoring Severe Weather Events and Climate*. Ed. by J. Jones, G. Guerova, J. Douša, G. Dick, S. de Haan, E. Pottiaux, O. Bock, R. Pacione, and R. van Malderen. Cham: Springer International Publishing, pp. 267–402. ISBN: 978-3-030-13901-8.

Boehm, J., A. E. Niell, P. Tregoning, and H. Schuh (2006a). “The Global Mapping Function (GMF) : A new empirical mapping function based on numerical weather model data”. In: *Geophys. Res. Lett.* 33, p. L07304. DOI: 10.1029/2005GL025546.

Boehm, J., B. Werl, and H. Schuh (2006b). “Troposphere mapping functions for GPS and very long baseline interferometry from European Centre for Medium-Range Weather Forecasts operational analysis data”. In: *J. Geophys. Res.* 11, pp. 2406–+. DOI: 10.1029/2005JB003629.

Bony, S. et al. (Aug. 2006). “How Well Do We Understand and Evaluate Climate Change Feedback Processes?” In: *Journal of Climate* 19.15, pp. 3445–3482. DOI: 10.1175/jcli3819.1. URL: https://doi.org/10.1175/jcli3819.1.

Boucher, O., J. Servonnat, A. L. Albright, O. Aumont, Y. Balkanski, and V. e. a. Bastrikov (2020). “Presentation and Evaluation of the IPSL-CM6A-LR Climate Model”. In: *Journal of Advances in Modeling Earth Systems* 12.7. e2019MS002010 10.1029/2019MS002010, e2019MS002010. DOI: https://doi.org/10.1029/2019MS002010. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.

1029/2019MS002010. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002010.

Box, G. E. P., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2016). *Time series analysis, forecasting and control*. Wiley.

Breiman, L. (1996). "Bagging Predictors". In: *Machine Learning* 24, pp. 123–140. URL: https://api.semanticscholar.org/CorpusID:207738357.

— (2001). "Random Forests". In: *Machine Learning* 45, pp. 5–32.

Breiman, L, J Friedman, R Olshen, and C Stone (1984). "Cart". In: *Classification and Regression Trees*.

Brockwell, P. J., R. A. Davis, P. J. Brockwell, and R. A. Davis (1991). "Stationary time series". In: *Time Series: Theory and Methods*, pp. 1–41.

Byun, S. H. and Y. E. Bar-Sever (2009). "A new type of troposphere zenith path delay product of the international GNSS service". In: *J. Geod.* 83.3, pp. 1–7. ISSN: 1432-1394. DOI: 10.1007/s00190-008-0288-8.

Caussinus, H. and O. Mestre (2004). "Detection and correction of artificial shifts in climate series". In: *J. R. Stat. Soc. : Ser. C (Appl. Stat.)* 53.3, pp. 405–425. DOI: 10.1111/j.1467-9876.2004.05155.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2004.05155.x.

Chakar, S., E Lebarbier, C. Lévy-Leduc, and S. Robin (2017). "A robust approach for estimating change-points in the mean of an AR(1) process". In: *Bernoulli* 23.2, pp. 1408–1447. ISSN: 13507265. URL: http://www.jstor.org/stable/44245623 (visited on 09/27/2023).

Colman, R. and B. J. Soden (2021). "Water vapor and lapse rate feedbacks in the climate system". In: *Rev. Mod. Phys.* 93 (4), p. 045002. DOI: 10.1103/RevModPhys.93.045002. URL: https://link.aps.org/doi/10.1103/RevModPhys.93.045002.

Costa, A. C. and A. Soares (2009). "Homogenization of Climate Data: Review and New Perspectives Using Geostatistics". In: *Mathematical Geosciences* 41.3, pp. 291–305. ISSN: 1874-8953. DOI: 10.1007/s11004-008-9203-3. URL: https://doi.org/10.1007/s11004-008-9203-3.

Cover, T. M. and P. E. Hart (1967). "Nearest neighbor pattern classification". In: *IEEE Trans. Inf. Theory* 13, pp. 21–27. URL: https://api.semanticscholar.org/CorpusID:5246200.

Dach, R., S. Lutz, P. Walser, and P. Fridez (2015). *Bernese GNSS Software Version 5.2. User manual*. DOI: 10.7892/boris.72297.

Dach, R., S. Schaer, D. Arnold, E. Orliac, L. Prange, A. Susnik, A. Villiger, and A. Jäggi (2018). *CODE final product series for the IGS*. DOI: 10.7892/boris.75876.3. URL: http://www.aiub.unibe.ch/download/CODE.

Davis, J. L., T. A. Herring, I. I. Shapiro, A. E. E. Rogers, and G. Elgered (1985). "Geodesy by radio interferometry : Effects of atmospheric modeling errors on estimates of baseline length". In: *Radio Science* 20, pp. 1593–1607.

Davis, J. L., G. Elgered, A. E. Niell, and C. E. Kuehn (1993). "Ground-based measurement of gradients in the "wet" radio refractivity of air". In: *Radio Science* 28, pp. 1003–1018. DOI: 10.1029/93RS01917. URL: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/93RS01917.

Dee, D. P. et al. (2011). "The ERA-Interim reanalysis: Configuration and performance of the data assimilation system". In: *Q. J. R. Meteorol. Soc.* 137.656, pp. 553–597. DOI: https://doi.org/10.1002/qj.828. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.828. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.828.

Domonkos, P. (2011). "Adapted Caussinus-Mestre Algorithm for Networks of Temperature series (ACMANT)". In: *Int. J. Geosci.* 02.03, pp. 293–309. DOI: 10.4236/ijg.2011.23032. URL: https://doi.org/10.4236/ijg.2011.23032.

— (Sept. 2021). "Combination of Using Pairwise Comparisons and Composite Reference Series: A New Approach in the Homogenization of Climatic Time Series with ACMANT". In: *Atmosphere* 12.9, p. 1134. DOI: 10.3390/atmos12091134. URL: https://doi.org/10.3390/atmos12091134.

Domonkos, P., J. A. Guijarro, V. Venema, M. Brunet, and J. Sigró (Apr. 2021). "Efficiency of Time Series Homogenization: Method Comparison with 12 Monthly Temperature Test Datasets". In: *Journal of Climate* 34.8, pp. 2877–2891. DOI: 10.1175/jcli-d-20-0611.1. URL: https://doi.org/10.1175/jcli-d-20-0611.1.

Douville, H. et al. (2021). "Water Cycle Changes". In: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 1055–1210. DOI: 10.1017/9781009157896.010.

Douville, H., S. Qasmi, A. Ribes, and O. Bock (2022). "Global warming at near-constant tropospheric relative humidity is supported by observations". In: *Communications Earth & Environment* 3. URL: https://doi.org/10.1038/s43247-022-00561-z.

Douville, H. and K. M. Willett (July 2023). "A drier than expected future, supported by near-surface relative humidity observations". In: *Science Advances* 9.30. DOI: 10.1126/sciadv.ade6253. URL: https://doi.org/10.1126/sciadv.ade6253.

Dunn, R. J. H., F. Aldred, N. Gobron, J. B. Miller, and K. M. Willett (2021). "Global Climate". In: *State of the Climate in 2020*. Vol. 102. 8. Bulletin of the American Meteorological Society, S11–S141. DOI: 10.1175/BAMS-D-21-0098.1. URL: https://doi.org/10.1175/BAMS-D-21-0098.1.

Dunn, R. J. H., M. J. B, W. K. M., and G. N. (Sept. 2023). "Global Climate". In: *State of the Climate in 2022*. Vol. 104. 9. Bulletin of the American Meteorological Society, S11–S145. DOI: 10.1175/bams-d-23-0090.1. URL: https://doi.org/10.1175/bams-d-23-0090.1.

Easterling, D. and T. Peterson (1995). "A new method for detecting undocumented discontinuities in climatological time series." In: *Int. J. Climatol.* 15, pp. 369–377. DOI: 10.1002/joc.3370150403.

Efron, B. and R. Tibshirani (1997). "Improvements on Cross-Validation: The 632+ Bootstrap Method". In: *Journal of the American Statistical Association* 92, pp. 548–560. URL: https://api.semanticscholar.org/CorpusID:18745711.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. Boca Raton, Florida, USA: Chapman & Hall/CRC.

Fact Sheet: Reanalysis (2023). [Online; accessed 2023-10-06]. URL: https://www.ecmwf.int/en/about/media-centre/focus/2023/fact-sheet-reanalysis.

Fisher, R. A. (1936). "The use of multiple measurements in taxonomic problems". In: *Annals of eugenics* 7.2, pp. 179–188.

Fix, E. and J. L. Hodges (1989). "Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties". In: *International Statistical Review* 57, p. 238. URL: https://api.semanticscholar.org/CorpusID:120323383.

Flato, G. et al. (2013). "Evaluation of climate models". In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]* Cambridge, UK: Cambridge University Press, pp. 741–882. DOI: 10.1017/CBO9781107415324.020.

Fu, W. J., R. J. Carroll, and S. Wang (2005). "Estimating misclassification error with small samples via bootstrap cross-validation". In: *Bioinformatics* 219, pp. 1979–86.

Granger, C. W. J. and M. J. Morris (1976). "Time Series Modelling and Interpretation". In: *Journal of the Royal Statistical Society*.

Guerova, G. et al. (2016). "Review of the state of the art and future prospects of the ground-based GNSS meteorology in Europe". In: *Atmospheric Measurement Techniques* 9.11, pp. 5385–5406. DOI: 10.5194/amt-9-5385-2016. URL: http://www.atmos-meas-tech.net/9/5385/2016/.

Guijarro, J. A. (2011). *User's guide to climatol. An R contributed package for homogenization of climatological series*. URL: http://webs.ono.com/climatol/climatol.html.

Gulev et al. (July 2021). "Changing State of the Climate System". In: *Climate Change 2021 – The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)].* Cambridge University Press, pp. 287–422. DOI: 10.1017/9781009157896.004. URL: https://doi.org/10.1017/9781009157896.004.

Hartmann, D. L. et al. (2013). "Observations: Atmosphere and surface". In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]* Cambridge, UK: Cambridge University Press. DOI: 10.1017/CBO9781107415324.008.

He, H. and E. A. Garcia (2009). "Learning from Imbalanced Data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9, pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.

Held, I. M. and B. J. Soden (2000). "Water Vapor Feedback and Global Warming". In: *Annual Review of Energy and the Environment* 25.1, pp. 441–475. DOI: 10.1146/annurev.energy.25.1.441. URL: https://doi.org/10.1146/annurev.energy.25.1.441.

— (2006). "Robust Responses of the Hydrological Cycle to Global Warming". In: *Journal of Climate* 19.21, pp. 5686 –5699. DOI: https://doi.org/10.1175/JCLI3990.1. URL: https://journals.ametsoc.org/view/journals/clim/19/21/jcli3990.1.xml.

Herring, T., R. King, M. Floyd, and S. McClusky (2015). "GAMIT Reference Manual. GPS Analysis at MIT GLOBK, Release 10. 6". In: *Massachusetts Institute of Technology*.

Hersbach, H. et al. (2020). "The ERA5 Global Reanalysis". In: *Q. J. R. Meteorol. Soc.* Accepted Author Manuscript.n/a. DOI: 10.1002/qj.3803. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803.

Hersbach, H. (2019). "Global reanalysis: goodbye ERA-Interim, hello ERA5". In: *ECMWF newsletter* 159, p. 17.

Ho, T. K. (1995). "Random decision forests". In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE, pp. 278–282.

Hofmann-Wellenhof, B., H. Lichtenegger, and E. Wasle (2007). *GNSS – Global Navigation Satellite Systems*. Springer Vienna. DOI: 10.1007/978-3-211-73017-1. URL: https://doi.org/10.1007/978-3-211-73017-1.

Hyndman, R. J. and Y. Khandakar (2008). "Automatic Time Series Forecasting: The forecast Package for R". In: *Journal of Statistical Software* 27, pp. 1–22.

Hyndman, R. J. et al. (2018). "forecast: Forecasting functions for time series and linear models". In: URL: https://api.semanticscholar.org/CorpusID:134205745.

IBM (2023). *K-Nearest Neighbors (KNN)*. Accessed on: 2023-08-31. URL: https://www.ibm.com/topics/knn.

Jones, P. D., S. C. B. Raper, R. S. Bradley, H. F. Diaz, P. M. Kellyo, and T. M. L. Wigley (1986). "Northern Hemisphere Surface Air Temperature Variations: 1851–1984". In: *J. Clim. Appl. Meteorol.* 25.2, pp. 161–179. DOI: 10.1175/1520-0450(1986)025<0161:NHSATV>2.0.CO;2.

José C. Pinheiro, D. M. B. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer New York, NY. DOI: 10.1007/b98882. URL: https://doi.org/10.1007/b98882.

Koreisha, S. G. and T. M. Pukkila (1995). "A Comparison Between Different Order-Determination Criteria for Identification of ARIMA Models". In: *Journal of Business & Economic Statistics* 13, pp. 127–131. URL: https://api.semanticscholar.org/CorpusID:155081484.

Kozubek, M., P. Krizan, and J. Lastovicka (Feb. 2020). "Homogeneity of the Temperature Data Series from ERA5 and MERRA2 and Temperature Trends". In: *Atmosphere* 11.3, p. 235. DOI: 10.3390/atmos11030235. URL: https://doi.org/10.3390/atmos11030235.

Lai, Y. and D. A. Dzombak (2020). "Use of the Autoregressive Integrated Moving Average (ARIMA) Model to Forecast Near-Term Regional Temperature and Precipitation". In: *Weather and Forecasting* 35.3, pp. 959–976. DOI: https://doi.org/10.1175/WAF-D-19-0158.1. URL: https://journals.ametsoc.org/view/journals/wefo/35/3/waf-d-19-0158.1.xml.

Lavielle, M. (2005). "Using penalized contrasts for the change-point problem". In: *Signal Process.* 85, pp. 1501–1510. URL: https://api.semanticscholar.org/CorpusID:478027.

Lebarbier, E. (2005). "Detecting Multiple Change-Points in the Mean of Gaussian Process by Model Selection". In: *Signal Processing* 85, pp. 717–736.

Lu, Q., R. Lund, and T. C. M. Lee (Mar. 2010). "An MDL approach to the climate segmentation problem". In: *The Annals of Applied Statistics* 4.1. DOI: 10.1214/09-aoas289. URL: https://doi.org/10.1214/09-aoas289.

Lund, R., H. Hurd, P. Bloomfield, and R. Smith (Nov. 1995). "Climatological Time Series with Periodic Correlation". In: *Journal of Climate* 8.11, pp. 2787–2809. DOI: 10.1175/1520-0442(1995)008<2787:ctswpc>2.0.co;2. URL: https://doi.org/10.1175/1520-0442(1995)008<2787:ctswpc>2.0.co;2.

Manabe, S. and R. T. Wetherald (1967). "Thermal Equilibrium of the Atmosphere with a Given Distribution of Relative Humidity". In: *Journal of Atmospheric Sciences* 24.3, pp. 241 –259. DOI: https://doi.org/10.1175/1520-0469(1967)024<0241:TEOTAW>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/atsc/24/3/1520-0469_1967_024_0241_teotaw_2_0_co_2.xml.

Menne, M. J. and C. N. Williams (2005). "Detection of Undocumented Changepoints Using Multiple Test Statistics and Composite Reference Series". In: *J. Clim.* 18.20, pp. 4271–4286. DOI: 10.1175/JCLI3524.1.

— (Apr. 2009). "Homogenization of Temperature Series via Pairwise Comparisons". In: *Journal of Climate* 22.7, pp. 1700–1717. DOI: 10.1175/2008jcli2263.1. URL: https://doi.org/10.1175/2008jcli2263.1.

Menne, M. J., C. N. Williams, and R. S. Vose (July 2009). "The U.S. Historical Climatology Network Monthly Temperature Data, Version 2". In: *Bulletin of the American Meteorological Society* 90.7, pp. 993–1008. DOI: 10.1175/2008bams2613.1. URL: https://doi.org/10.1175/2008bams2613.1.

Mestre, O. et al. (Jan. 2013). "HOMER: A homogenization software - methods and applications". In: *Idojaras* 117.

Mitchell, T. D. and P. D. Jones (2005). "An improved method of constructing a database of monthly climate observations and associated high-resolution grids". In: *International Journal of Climatology* 25.6, pp. 693–712. DOI: 10.1002/joc.1181. URL: https://doi.org/10.1002/joc.1181.

Newey, W. and K. D. West (1986). "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation-consistent Covariance Matrix". In: *Econometrics eJournal*.

Nguyen, K. N., A. Quarello, O. Bock, and E. Lebarbier (2021). "Sensitivity of Change-Point Detection and Trend Estimates to GNSS IWV Time Series Properties". In: *Atmosphere* 12.9, p. 1102. DOI: 10.3390/atmos12091102. URL: https://doi.org/10.3390/atmos12091102.

Niell, A. E. (1996). "Global mapping functions for the atmosphere delay at radio wavelengths". In: *J. Geophys. Res. : Solid Earth* 101.B2, pp. 3227–3246. DOI: 10.1029/95jb03048. URL: https://doi.org/10.1029/95jb03048.

— (2000). "Improved atmospheric mapping functions for VLBI and GPS". In: *Earth, Planets and Space* 52, pp. 699–702. DOI: 10.1186/BF03352267.

Ning, T., J. Wickert, Z. Deng, S. Heise, G. Dick, S. Vey, and T. Schöne (2016). "Homogenized Time Series of the Atmospheric Water Vapor Content Obtained from the GNSS Reprocessed Data". In: *J. Clim.* 29.7, pp. 2443–2456. DOI: 10.1175/JCLI-D-15-0158.1.

Parracho, A. C., O. Bock, and S. Bastin (2018). "Global IWV trends and variability in atmospheric reanalyses and GPS observations". In: *Atmos. Chem. Phys.* 18.22, pp. 16213–16237. DOI: 10.5194/acp-18-16213-2018. URL: https://www.atmos-chem-phys.net/18/16213/2018/.

Parracho, A. C. B. (Dec. 2017). "Study of trends and variability of atmospheric water vapour with climate models and observations from global GNSS network". Theses. Université Pierre et Marie Curie - Paris VI. URL: https://theses.hal.science/tel-01881083.

Peixoto, J. and A. H. Oort (Dec. 1996). "The Climatology of Relative Humidity in the Atmosphere". In: *Journal of Climate* 9.12, pp. 3443–3463. DOI: 10.1175/1520-0442(1996)009<3443:tcorhi>2.0.co;2. URL: https://doi.org/10.1175/1520-0442(1996)009<3443:tcorhi>2.0.co;2.

Peterson, T. C., R. Vose, R. Schmoyer, and V. Razuvaëv (Sept. 1998a). "Global historical climatology network (GHCN) quality control of monthly temperature data". In: *International Journal of Climatology* 18.11, pp. 1169–1179. DOI: 10.1002/(sici)1097-0088(199809)18:11<1169::aid-joc309>3.0.co;2-u. URL: https://doi.org/10.1002/(sici)1097-0088(199809)18:11<1169::aid-joc309>3.0.co;2-u.

Peterson, T. C. et al. (1998b). "Homogeneity adjustments of in **situ atmospheric** climate data: A review". In: *Int. J. Climatol. : A J. R. Meteorol. Soc.* 18.13, pp. 1493–1517.

Quarello, A. (Dec. 2020). "Development of new homogenisation methods for GNSS atmospheric data. Application to the analysis of climate trends and variability". PhD thesis. Sorbonne Universite ; IGN (Institut National de l'Information Géographique et Forestière); https://hal.archives-ouvertes.fr/tel-03118629. URL: https://hal.archives-ouvertes.fr/tel-03118629.

Quarello, A., O. Bock, and E. Lebarbier (2022). "GNSSseg, a Statistical Method for the Segmentation of Daily GNSS IWV Time Series". In: *Remote Sensing* 14.14, p. 3379. DOI: 10.3390/rs14143379. URL: https://doi.org/10.3390/rs14143379.

Radke, N., K. Keller, R. Yousefpour, and M. Hanewinkel (2020). "Identifying decision-relevant uncertainties for dynamic adaptive forest management under climate change". In: *Climatic Change* 163, pp. 891 –911. URL: https://api.semanticscholar.org/CorpusID:226286670.

Rebischung, P. and R. Schmid (Dec. 2016). "IGS14/igs14.atx: a new Framework for the IGS Products". In: *AGU Fall Meeting Abstracts*. Vol. 2016, G41A-0998, G41A–0998.

Reeves, J., J. Chen, X. L. Wang, R. Lund, and Q. Q. Lu (2007). "A Review and Comparison of Changepoint Detection Techniques for Climate Data". In: *J. Appl. Meteorol. Climatol.* 46.6, pp. 900–915. DOI: 10.1175/JAM2493.1. eprint: https://doi.org/10.1175/JAM2493.1. URL: https://doi.org/10.1175/JAM2493.1.

Rienecker, M. M. et al. (2011). "MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications". In: *Journal of Climate* 24.14, pp. 3624 –3648. DOI: https://doi.org/10.1175/JCLI-D-11-00015.1. URL: https://journals.ametsoc.org/view/journals/clim/24/14/jcli-d-11-00015.1.xml.

Rodriguez, J. D., A. Perez, and J. A. Lozano (2010). "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.3, pp. 569–575. DOI: 10.1109/TPAMI.2009.187.

Saastamoinen, J. (Mar. 1973). "Atmospheric Correction for the Troposphere and Stratosphere in Radio Ranging Satellites". In: *The Use of Artificial Satellites for Geodesy*. American Geophysical Union, pp. 247–251. DOI: 10.1029/gm015p0247. URL: https://doi.org/10.1029/gm015p0247.

Santer, B. D. et al. (2021). "Using Climate Model Simulations to Constrain Observations". In: *Journal of Climate*. URL: https://api.semanticscholar.org/CorpusID:236407107.

Schneider, T., P. A. O'Gorman, and X. J. Levine (July 2010). "WATER VAPOR AND THE DYNAMICS OF CLIMATE CHANGES". In: *Reviews of Geophysics* 48.3. DOI: 10.1029/2009rg000302. URL: https://doi.org/10.1029/2009rg000302.

Schroeder, M., M. Lockhoff, J. Forsythe, H. Cronk, T. V. Haar, and R. Bennartz (2016). "The GEWEX Water Vapor Assessment: Results from Intercomparison, Trend, and Homogeneity Analysis of Total Column Water Vapor". In: *J. Appl. Meteorol. Climatol.* 55.7, pp. 1633–1649. DOI: 10.1175/jamc-d-15-0304.1. URL: https://doi.org/10.1175/jamc-d-15-0304.1.

Shaub, D. (2020). "Fast and accurate yearly time series forecasting with forecast combinations". In: *International Journal of Forecasting* 36.1. M4 Competition, pp. 116–120. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2019.03.032. URL: https://www.sciencedirect.com/science/article/pii/S0169207019301566.

Sherwood, S. C., R. Roca, T. M. Weckwerth, and N. G. Andronova (Apr. 2010). "Tropospheric water vapor, convection, and climate". In: *Reviews of Geophysics* 48.2. DOI: 10.1029/2009rg000301. URL: https://doi.org/10.1029/2009rg000301.

Shibata, R. (1976). "Selection of the order of an autoregressive model by Akaike's information criterion". In: *Biometrika* 63, pp. 117–126.

Shumway, R. H. and D. S. Stoffer (2011). *Time Series Analysis and Its Applications*. Springer International Publishing. DOI: 10.1007/978-3-319-52452-8. URL: https://doi.org/10.1007/978-3-319-52452-8.

— (2017). *Time Series Analysis and Its Applications. Fourth Edition.* Springer International Publishing. DOI: 10.1007/978-3-319-52452-8. URL: https://doi.org/10.1007/978-3-319-52452-8.

Sterl, A. (Oct. 2004). "On the (In)Homogeneity of Reanalysis Products". In: *Journal of Climate - J CLIMATE* 17, pp. 3866–3873. DOI: 10.1175/1520-0442(2004)017<3866:OTIORP>2.0.CO;2.

Szentimrey, T. (2008). "Development of MASH homogenization procedure for daily data. Proceedings of the fifth seminar for homogenization and quality control in climatological databases". In: *WCDMP-No. 71*, 123–130. URL: http://www.wmo.int/pages/prog/wcp/wcdmp/documents/WCDMP71.pdf.

Teunissen, P. J. G. and O. Montenbruck (2017a). "Springer Handbook of Global Navigation Satellite Systems". In: *Springer Handbook of Global Navigation Satellite Systems*. DOI: https://doi.org/10.1007/978-3-319-42928-1.

Teunissen, P. J. and O. Montenbruck, eds. (2017b). *Springer Handbook of Global Navigation Satellite Systems*. Springer International Publishing. DOI: 10.1007/978-3-319-42928-1. URL: https://doi.org/10.1007/978-3-319-42928-1.

Thorne, P. W. and R. S. Vose (2010). "Reanalyses Suitable for Characterizing Long-Term Trends". In: *Bulletin of the American Meteorological Society* 91.3, pp. 353 –362. DOI: https://doi.org/10.1175/2009BAMS2858.1. URL: https://journals.ametsoc.org/view/journals/bams/91/3/2009bams2858_1.xml.

TIBCO (2023). *What is a Random Forest?* Accessed on: 2023-08-31. URL: https://www.tibco.com/reference-center/what-is-a-random-forest.

Trenberth et al. (2013). "Observations: Surface and Atmospheric Climate Change". In: *In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)].* Cambridge, UK: Cambridge University Press. DOI: 10.1017/CBO9781107415324.008.

Trenberth, K. E. (1998). "Atmospheric Moisture Residence Times and Cycling: Implications for Rainfall Rates and Climate Change". In: *Climatic Change* 39.4, pp. 667–694. DOI: 10.1023/a:1005319109110. URL: https://doi.org/10.1023/a:1005319109110.

Van Malderen, R. et al. (2020). "Homogenizing GPS Integrated Water Vapor Time Series: Benchmarking Break Detection Methods on Synthetic Data Sets". In: *Earth Space Sci.* 7.5. e2020EA001121 2020EA001121, e2020EA001121. DOI: 10.1029/2020EA001121. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020EA001121. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020EA001121.

Van Malderen, R., H. Brenot, E. Pottiaux, S Beirle, C. Hermans, M. De Maziere, T Wagner, H. De Backer, and C. Bruyninx (Aug. 2014). "A multi-site intercomparison of integrated water vapour observations for climate change analysis". In: *Atmos. Meas. Tech.* 7. DOI: 10.5194/amt-7-2487-2014.

Venema, V. K. C. et al. (2012). "Benchmarking homogenization algorithms for monthly data". In: *Clim. Past* 8.1, pp. 89–115. DOI: 10.5194/cp-8-89-2012. URL: https://www.clim-past.net/8/89/2012/.

Vey, S., R. Dietrich, M. Fritsche, A. Rülke, P. Steigenberger, and M. Rothacher (2009). "On the homogeneity and interpretation of precipitable water time series derived from global GPS observations". In: *J. Geophys. Res. : Atmos.* 114.D10.

Wang, J., A. Dai, and C. Mears (2016). "Global water vapor trend from 1988 to 2011 and its diurnal asymmetry based on GPS, radiosonde, and microwave satellite measurements". In: *Journal of Climate* 29.14, pp. 5205–5222.

Wang, X. L., H. Chen, Y. Wu, Y. Feng, and Q. Pu (2010). "New Techniques for the Detection and Adjustment of Shifts in Daily Precipitation Data Series". In: *Journal of Applied Meteorology and Climatology* 49.12, pp. 2416 –2436. DOI: https://doi.org/10.1175/2010JAMC2376.1. URL: https://journals.ametsoc.org/view/journals/apme/49/12/2010jamc2376.1.xml.

White, H. L. (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity". In: *Econometrica* 48, pp. 817–838.

Wikipedia contributors (2023). *Decision tree learning — Wikipedia, The Free Encyclopedia*. [Online; accessed 30-August-2023]. URL: https://en.wikipedia.org/w/index.php?title=Decision_tree_learning&oldid=1169691500.

Yan, B., R. Mu, J. Guo, Y. Liu, J. Tang, and H. Wang (2022). "Flood risk analysis of reservoirs based on full-series ARIMA model under climate change". In: *Journal of Hydrology* 610, p. 127979. ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2022.127979. URL: https://www.sciencedirect.com/science/article/pii/S0022169422005546.

Yohai, V. J. and R. H. Zamar (1988). "High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale". In: *Journal of the American Statistical Association* 83, pp. 406–413.

Zeileis, A. (2006). "Object-oriented Computation of Sandwich Estimators". In: *Journal of Statistical Software* 16, pp. 1–16.

Zhang, N. R. and D. O. Siegmund (2007). "A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data". In: *Biometrics* 63. URL: https://api.semanticscholar.org/CorpusID:42771901.

Zhu, S. Y. and E. Groten (1988). "Relativistic effects in GPS". In: *GPS-Techniques Applied to Geodesy and Surveying*. Springer-Verlag, pp. 41–46. DOI: 10.1007/bfb0011322. URL: https://doi.org/10.1007/bfb0011322.

Zumberge, J. F., M. B. Heflin, D. C. Jefferson, and M. M. Watkins (1997). "Precise point positioning for the efficient and robust analysis of GPS data from large networks". In: *J. Geophys. Res.* 102, pp. 5005–5017.

Zwiers, F. W. and H. von Storch (1995). "Taking Serial Correlation into Account in Tests of the Mean." In: *Journal of Climate* 8, pp. 336–351.