Déformation des cycles saisonniers de variables climatiques

Alix Rigal

▶ **To cite this version:**

Alix Rigal. Déformation des cycles saisonniers de variables climatiques. Climatologie. Université Paul Sabatier - Toulouse III, 2020. Français. NNT : 2020TOU30112 . tel-03118539

HAL Id: tel-03118539
https://theses.hal.science/tel-03118539

Submitted on 22 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *23 juin 2020* **par :**

Alix Rigal

Déformation des cycles saisonniers de variables climatiques

JURY

Marc Lavielle	Dir. de Rech., CNRS CMAP	Rapporteur
David B. Stephenson	Professeur, Univ. of Exeter	Rapporteur
Valérie Monbet	Professeur, Univ. de Rennes	Examinatrice
Sylvie Parey	Chercheur senior, EDF R&D	Examinatrice
Pascal Yiou	Chercheur senior LSCE/IPSL	Examinateur
Jean-Marc Azaïs	Professeur, Univ. P. Sabatier	Directeur de Thèse
Aurélien Ribes	Chercheur, CNRM-GAME	Directeur de Thèse
Thierry Klein	Professeur, ENAC	Président du Jury

École doctorale et spécialité :
MITT : Domaine STIC : Réseaux, Télécoms, Systèmes et Architecture

Unité de Recherche :
Centre National de Recherches Météorologiques (UMR 3589)
Institut de Mathématiques de Toulouse (UMR 5219)

Directeur(s) de Thèse :
Jean-Marc Azaïs et *Aurélien Ribes*

Rapporteurs :
Marc Lavielle et *David B. Stephenson*

Remerciements

Je ne serai pas arrivé au bout de ce travail de recherche sans la collaboration de plusieurs personnes. Je tiens à vous remercier à travers ces quelques lignes, pour m'avoir accompagné dans ce long voyage qu'est la thèse. D'avance, je souhaite présenter mes excuses à ceux que j'aurais malencontreusement oubliés.

Tout d'abord je tiens à exprimer ma gratitude envers mes rapporteurs Marc Lavielle et David Stephenson pour leur relecture en profondeur de mon manuscrit ainsi que leurs commentaires constructifs. Plus largement, je suis redevable à mon jury de thèse qui a su être disponible, patient et compréhensif dans cette période sanitaire inédite. En particulier Thierry Klein, qui a bien voulu se prêter au jeu difficile de président de jury pour une soutenance dans des conditions extraordinaires, ainsi que pour sa lecture attentive de ce document. Je remercie également Sylvie Parey pour ses corrections détaillée et ses contributions à mon manuscrit.

Un aspect important de cette expérience est qu'elle m'a permis d'apprendre et découvrir des domaines qui, pour moi, étaient nouveaux. J'ai notamment apprécié le groupe de travail 'biblioStat' qui a eu pour moi une impacte très positif. Je remercie à cet égard l'ensemble des participants réguliers (ils se reconnaîtrons) et occasionnels pour leurs retours et exposés. Plus particulièrement, je remercie Thomas Rieutord sans qui il aurait été impossible de maintenir un rythme hebdomadaire. Dans ce même domaine de recherche, les séminaires de l'IMT m'ont aussi été d'une grande aide. Je remercie également Julien Cattiaux pour m'avoir fait découvrir le système climatique ainsi que l'équipe AMACS pour m'avoir accueilli et fait approfondir ces notions au travers des fameux 'jeudi du climat'.

J'ai eu la chance durant mon doctorat de pouvoir m'investir dans les enseignements de l'ENM sous la direction de William Ohayon. Ce fut un plaisir de faire mes premier pas avec toi et j'ai toujours apprécié nos échanges. Je remercie également Pascal Laveau, Marie-Pierre Traullé et David Pollack ainsi que toute l'équipe pédagogique pour m'avoir accueilli aussi chaleureusement.

Mon travail de recherche sur le terrain a été facilité par les laboratoires m'ayant accueilli durant ces quatres années de thèse. Je tiens donc en premier lieu à remercier tous les membres du CNRM et de l'IMT, qui ont répondu avec calme et patience aux questions quotidiennes dont je les accablais. Je remercie également toute l'équipe DEVI et plus largement le département SINA pour leurs accueil à l'ENAC durant ces derniers mois d'écriture, de soutenance et de confinement. J'adresse mes plus sincères remerciements aux personnels administratifs de Météo-France et de l'IMT pour les nombreuses fois où j'ai fait appel à eux, tout particulièrement Agnès Requis.

La thèse est un travail de longue haleine qui aurait pu être plus épuisant qu'enrichissant sans les soutiens dont j'ai pu profité. Je tiens tout d'abord à remercier Jean-Marc Tregan dont l'amitié a su me motiver en me partageant ses problèmes de mathématiques. Je remercie toutes les personnes avec qui j'ai partagé mes études et notamment ces années de thèse. Enfin, toutes ces bonnes choses n'auraient su être magnifiées sans l'appui sans faille de ma compagne Edith et de mes proches.

Bien sûr à cette liste non-exhaustive manque les deux protagonistes principaux : mes directeurs de thèse Jean-Marc Azaïs et Aurélien Ribes. Mes derniers remerciements leurs sont évidement dédiés ; tout d'abord pour la confiance qu'ils m'ont accordée en acceptant d'encadrer ce travail doctoral, pour leurs multiples conseils et les nombreux encouragements qu'ils m'ont prodigués. Je les remercie aussi pour leurs disponibilité ainsi que toutes ces heures consacrées à diriger cette recherche. Enfin, j'ai été extrêmement sensible à leurs qualités humaines respectives d'écoute et de compréhension tout au long de ce travail.

Table des Matières

Table des notations et acronymes

OMM/WMO	Organisation Météorologique Mondiale/ *World Meteorological Organization*
NOAA	*National Oceanic and Atmospheric Administration*
IPCC/GIEC	*Intergovernmental Panel on Climate Change*/Groupe d'Experts Intergouvernemental sur l'Évolution du climat
DJU	Degré Jour Unifié
DJC	Degré Jour de Climatisation
ETP	Evapo-Transpiration Potentielle
CCNUCC	Convention-Cadre des Nations Unies sur les Changements Climatiques
EBMs	*Energy Balance Models*, Modèles de bilan énergétique
EMICs	*Earth-system Models of Intermediate Complexity*, Modèles Terre à Complexité Intermédiaire
GCMs	*General Circulation Models*, Modèles de Circulation Générale
AGCM	*Atmospheric General Circulation Model*, Modèle de Circulation Générale Atmosphérique
OGCM	*Oceanic General Circulation Model*, Modèle de Circulation Générale Océanique
AOGCM	*Atmosphere-Ocean Coupled General Circulation Models*, Modèles de Circulation Générale Océan-Atmosphère (couplés)
ESMs	*Earth System Models*, Modèles du Système Terre
CMIP	*Coupled Models Intercomparison Project*, Projet d'Intercomparaison des Modèles Couplés

WCRP	*World Climate Research Program*, Programme Mondial de la Recherche sur le Climat
RCP	*Representative Concentration Pathway*
PIOMAS	*Pan-Arctic Ice Ocean Modeling and Assimilation System*
SQR	Séries Quotidiennes de Référence
RKHS	*Reproducing Kernel Hilbert Space*, Espaces de Hilbert à noyaux reproduisant
FDR/CDF	Fonction de répartition/*Cumulative Distribution Function*
IID	Indépendant et Identiquement Distribué
INID	Indépendant et Non-Identiquement Distribué
GLM	*Generalized linear model*, Modèle Linéaire Généralisé

Notation générale

f, g, h	Les trois composantes du modèle multiplicatif, respectivement cycle annuel de référence, tendance annuelle et delta cycle
d, y	Représentent les jours et années
$T_{d,y}$	Température moyenne observée au jour d de l'année y
$Pr_{d,y}$	Précipitation observée au jour d de l'année y
I_n	Matrice identité de dimension n
$\mathbb{1}$	Fonction indicatrice
$[\![a, b]\!]$	Entiers consécutifs allant de a à b
$[x]$	Partie entière de x
λ	Paramètre de régularisation pour l'estimation des splines
AIC	*Akaïk Information Criterion*, critère d'information d'Akaike
BIC/SIC	*Bayesian Information Criterion*, Critère d'Information Bayésien (Schwarz)
C_p	Cp de Mallows
$\mathcal{U}[0, 1]$	Loi uniforme sur $[0, 1]$
$N(\mu, \sigma)$	Loi normale de moyenne μ et écart type σ
df	Degrés de liberté
$\|.\|_p$	Norme usuelle des espaces $\mathcal{L}^p$
$\mathbb{N}$	Ensemble des entiers naturels
$\mathbb{S}^1$	Cercle
$Card(E)$	Cardinal de l'ensemble **E**

Chapitre 2 :

Espaces de Hilbert à noyaux reproduisant

$\mathbf{X}$	Un ensemble arbitraire
$\mathcal{H}$	Un espace de Hilbert
$\langle , \rangle_{\mathcal{H}}$	Produit scalaire de $\mathcal{H}$
$\|f\|_{\mathcal{H}}$	Norme induite par $\mathcal{H}$
$\mathbb{R}^{\mathbf{X}}$	Fonctions à valeurs réelles sur $\mathbf{X}$
L_x	Fonctionnelle linéaire d'évaluation
$\mathcal{K}$	Noyau symétrique positif
$\bigoplus$	Somme directe de deux sous-espaces vectoriels
$\mathcal{H}^m$	Espaces de Sobolev
P^m	L'espace des polynômes de degrés au plus $m-1$
N	Matrice associée à la base spline naturelle
Ω_N	Matrice de Gram associée à la base spline
$\text{Vect}\{g_i, i \in [\![1, n]\!]\}$	Espace vectoriel engendré par les g_i
$\mathbf{E}^{\perp}$	Supplémentaire orthogonal de $\mathbf{E}$
$(t)_+ = \max(0, t)$	Partie positive de t
S_λ	Matrice de lissage spline

Régression quantile

$\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_k$	Variables aléatoires
$y, x_1, \ldots, x_k$	Réalisations des variables aléatoires $\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_k$
X	Matrice des régresseurs, *design matrix*
$F_{\mathbf{Y}}$	Fonction de répartition de $\mathbf{Y}$
$Q_{\mathbf{Y}}(\tau)$	Quantile τ de la variable aléatoire $\mathbf{Y}$
ρ_τ	Fonction coût (*hinge loss*) associée au quantile τ
$\hat{\beta}_\tau = \hat{\beta(\tau)}$	Paramètres estimés de la régression quantile pour le quantile τ
$\mathbb{1}$	Fonction indicatrice
$\hat{q}_\tau$	Quantile empirique
$\mathbf{1}$	Vecteur dont toutes les coordonnées valent 1
$N(\mu, \sigma)$	Loi normale de moyenne μ et écart type σ

Sélection de modèles et erreur de généralisation

$T = (X_1, Y_1), \ldots, (X_n, Y_n)$	Variables aléatoires représentant l'échantillon
$(x_1, y_1), \ldots, (x_n, y_n)$	Réalisation de T
$\hat{f}$	Estimateur entraîné sur T
L	Fonction coût
Err_T	Erreur de généralisation
Err	Erreur de test
$\bar{err}$	Erreur d'entraînement
$\mathbb{E}_Y[.]$	Espérance calculée uniquement sur la variable Y
CV	Validation croisée K-fold

Chapitre 3 :

Étude de l'espérance d'une variable climatique

WMO	Moyenne sur les 30 années passées remise à jour toutes les décennies
WMO reset	Moyenne sur les 30 années passées remise à jour tous les ans
OCN	Moyenne sur les 15 années passées remise à jour toutes les décennies
Hinge	Modèle linéaire produisant une ligne brisé réajustée toutes les décennies
Hinge reset	Modèle linéaire produisant une ligne brisée réajustée tous les ans
$\hat{f}, \hat{g}, \hat{h}$	Estimateur des composantes du modèle multiplicatif
df_f, df_g, df_h	Degrés de liberté de $\hat{f}, \hat{g}, \hat{h}$
MSE	Erreur quadratique moyenne
R^2	Coefficient de détermination
$PRESS$	Validation croisée 10-fold

Chapitre 4 :

Étude de la déformation de la distribution d'une variable

$\widehat{T}_{d,y,\tau}$	Estimation du quantile τ de température du jour d de l'année y
$\widehat{Pr}_{d,y,\tau}$	Estimation du quantile τ des précipitations du jour d de l'année y
$g(.)$	Spline sur les températures moyennes annuelles
$g(., \tau)$	Tendance annuelle pour chaque quantile τ
$f(d, \tau), h(d, \tau)$	Cycle de référence et delta cycle utilisés pour décrire le quantile τ
10-fold	Validation croisée K-fold avec $K = 10$
$\overline{Err}$	Moyenne sur les quantiles de la validation croisée 10-fold

Introduction

Les normales climatiques sont habituellement calculées comme des moyennes sur une période observée de 30 ans et remises à jour tous les dix ans. Durant ces dernières années, les différents organismes de suivi climatique ont eu l'occasion de présenter de nombreuses valeurs de température au-dessus des normales de saison.

C'est le cas, par exemple, pour les moyennes mensuelles de Toulouse (figure 1.1) de ces dix dernières années. La normale y est calculée à l'échelle mensuelle comme une moyenne évaluée pour chaque mois sur la période climatologique de référence, 1981-2010, comme préconisée par l'Organisation Météorologique Mondiale (OMM). On peut y voir, en plus du biais, un nombre d'anomalies chaudes (en rouge) déraisonnablement élevées (environ 3 fois plus) par rapport à la normale. Un constat similaire mais plus erratique peut être fait concernant les variables de

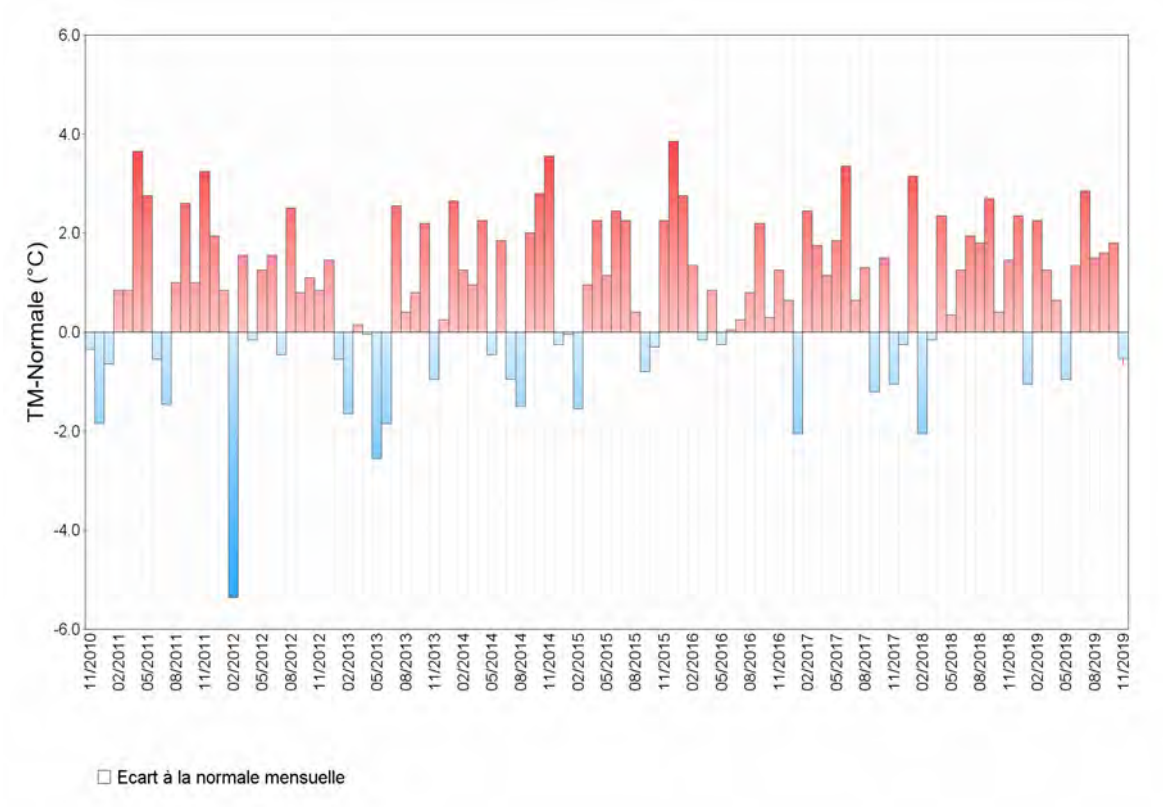


FIGURE 1.1 – Écart des températures moyennes mensuelles observées par rapport à la moyenne mensuelle de référence 1981-2010 de novembre 2010 à novembre 2019 sur la station de Toulouse-Blagnac.

températures moyennes journalière. La figure 1.2 montre une normale climatique décrite, pour le climat Toulousain de août 2016 à août 2019. En rouge, sont représentées les observations de températures moyennes se situant au-dessus de la normale (anomalies positives) et, en bleu, celles se situant en dessous (anomalies négatives). On constate que la proportion d'anomalies positives est en moyenne bien supérieure à celle d'anomalies négatives. Ce biais est également observable sur les séries d'anomalies de la précédente décennie, et semble montrer un retard de la normale sur l'état moyen du climat présent. Cela pose la question suivante : quelle

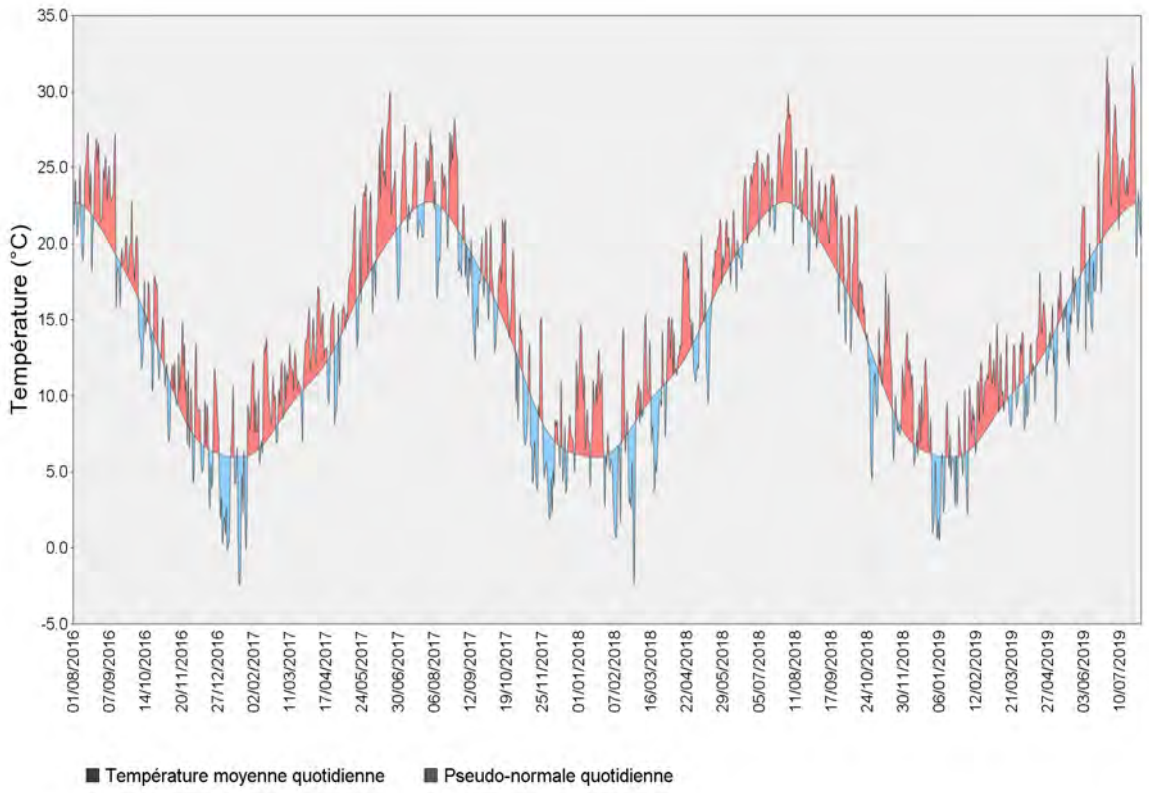


FIGURE 1.2 – Température moyenne journalière - 1^{er} août 2016 au 1^{er} août 2019 à Toulouse-Blagnac. La normale quotidienne y est représentée par la courbe noire, les anomalies positives (par rapport à cette normale) y sont représentées en rouge et les négatives en bleu.
Source : climascope.

référence représente une normale ? Cela a-t-il même un sens de la considérer dans un climat non-stationnaire ? Pour répondre à ces questions, il est nécessaire, dans un premier temps, de préciser les contours du système étudié ainsi que le sens donné au terme "climat".

1.1 Le système climatique

Avant d'introduire la notion de climat, il est nécessaire de définir proprement les contours du système étudié. Nous reproduisons ici la description fournie dans le dernier rapport du GIEC (Groupe d'Experts Intergouvernemental sur l'Évolution du Climat).

'The climate system is the highly complex system consisting of five major components : the atmosphere, the hydrosphere, the cryosphere, the lithosphere and the biosphere and the interactions between them. The climate system evolves in time under the influence of its own internal dynamics and because of external forcings such as volcanic eruptions, solar variations and anthropogenic forcings such as the changing composition of the atmosphere and land-use change.'
IPCC Special Report on Global Warming of 1.5 *°C, 2018, Glossary, p.545-546.*

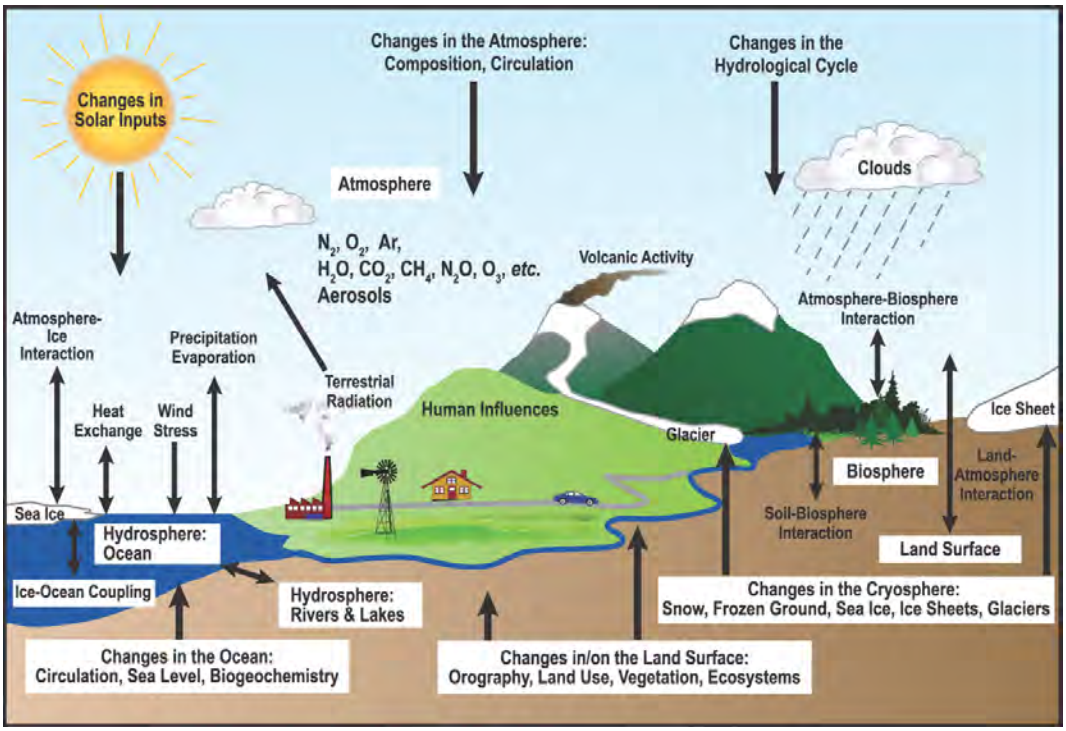


FIGURE 1.3 – Représentation schématique de chacune des composantes climatiques et de ses principales interactions.
Source : IPCC AR4 (2007) FAQ 1.2 Fig.1.

Il est intéressant de remarquer que cette définition donne, dans un premier temps, les objets constitutifs du système étudié, soit l'atmosphère, l'hydrosphère, la cryosphère, la lithosphère et la biosphère et, dans un deuxième temps, les interactions de ce système entre ses éléments constitutifs ou avec des éléments extérieurs (tels que le soleil et les volcans). Troisièmement, le système climatique est muni d'une chronologie, en effet, il évolue en fonction du temps sous l'influence de ses interactions (Figure 1.3). Dans la pratique, l'état du système désignera l'ensemble des variables climatiques pouvant permettre de le décrire (e.g. température, pression, albédo, etc.). Les interactions sont alors les dépendances entre une variable et toutes les autres, par exemple, par le biais d'équations différentielles.

1.2 La notion de climat

La notion de climat peut être précisée par comparaison à la météo à travers la célèbre formule *"Climate is what you expect, weather is what you get."*. Cette formule, popularisée en partie par Robert Heinlein, a le défaut d'être ambiguë. En effet, nous pourrions croire, à tort, que seule la valeur moyenne est d'intérêt en climat et qu'elle exclut de facto l'étude des valeurs extrêmes ou toutes autres parties de la distribution.

En réalité, cette version simplifiée trouve son origine dans un ouvrage de 1908 du géographe Andrew John Herbertson [58]. Celui-ci permet de préciser son sens premier : *'By climate we mean the average weather as ascertained by many years' observations. Climate also takes into account the extreme weather experienced during that period. Climate is what on an average we may expect, weather is what we actually get.'*

Autrement dit, pour Herbertson le climat est l'étude de la distribution des valeurs possibles d'une variable climatique sur une période temporelle donnée. Son étude nécessite l'intégration de plusieurs années d'observations. La météo, quant à elle, est une réalisation du climat, un tirage dans cette distribution de valeurs possibles. Par analogie avec un jeu de dés, l'objectif du climat serait de nous donner la probabilité de chaque face ainsi que la dépendance entre chaque lancé, celui de la météo serait de prévoir les lancés suivants. La définition donnée par le GIEC va dans ce sens et permet d'élargir encore un peu plus cette dernière définition.

"Climate in a narrow sense is usually defined as the average weather, or more rigorously, as the statistical description in terms of the mean and variability of relevant quantities over a period of time ranging from months to thousands or millions of years. The classical period for averaging these variables is 30 years, as defined by the World Meteorological Organization. The relevant quantities are most often surface variables such as temperature, precipitation and wind. Climate in a wider sense is the state, including a statistical description, of the climate system."
IPCC Special Report on Global Warming of 1.5 ^{o}*C, 2018, Glossary, p.544.*

Autrement dit, le climat est l'analyse de l'état du système climatique et, plus particulièrement, à l'aide de statistiques permettant de décrire, tout ou partie, de la distribution de probabilité de la quantité étudiée. Plus simplement, le climat est la distribution probabiliste des évènements météorologiques possibles. Par exemple, dans cette thèse nous nous intéresserons à l'évolution de la distribution des variables climatiques univariées, journalières à une localisation donnée, au cours du 20$^{\text{ème}}$ et du 21$^{\text{ème}}$ siècles.

1.3 Changement climatique

L'étude du climat est, en soit, une science extrêmement intéressante et complexe de par son interaction avec plusieurs disciplines - mathématiques, physique, mécanique, chimie et biologie entre autres. Néanmoins, elle n'aurait pas l'importance que l'on lui octroie aujourd'hui s'il n'était pas question des contraintes qu'un changement climatique aurait sur nos sociétés. Comme le montre la figure 1.4 issue du rapport du deuxième groupe de travail du GIEC, des tendances sur les dernières décennies de l'impact attribuable au changement climatique ont déjà pu être observées à travers le globe.

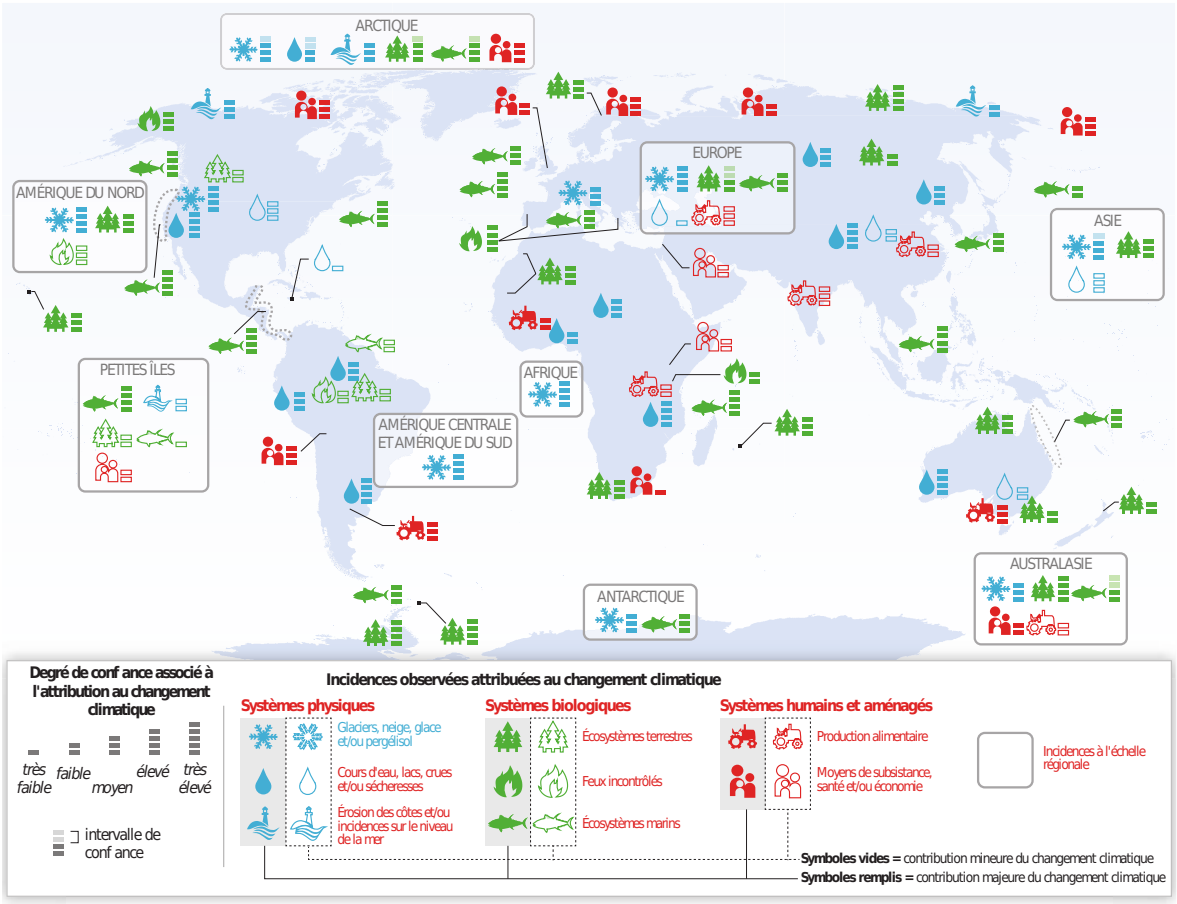


FIGURE 1.4 – Tendances mondiales des incidences attribuées au changement climatique observées au cours des dernières décennies, fondées sur les résultats des études réalisées depuis la publication du quatrième rapport d'évaluation du GIEC. Les symboles indiquent le type d'incidence, la contribution relative du changement climatique (majeure : symbole rempli ou mineure : symbole vide) aux incidences observées, et le degré de confiance correspondant indiqué par des barres horizontales. Source : Contribution du groupe de travail II (Incidences, adaptation et vulnérabilité) au $5^{\text{ème}}$ rapport d'évaluation du GIEC.

Les implications socio-économiques sont difficiles à prévoir avec précision à l'horizon 2100. Une majeure partie de ces incertitudes est liée à l'évolution des activités humaines, soit principalement aux scénarios d'émission de gaz à effet de serre, ainsi qu'à la vulnérabilité et la capacité d'adaptation des populations. Au cours de l'histoire de l'humanité, les peuples et les sociétés ont réussi, avec plus ou moins de succès, à s'adapter au climat. L'acquisition de cette forme de résilience à été possible par des variations lentes du climat, ce qui n'est pas le cas du changement en cours. Comme l'ensemble des variables climatiques peuvent avoir une incidence, cela motive le GIEC à avoir une définition précise du changement climatique.

Climate change refers to a change in the state of the climate that can be identified (e.g., by using statistical tests) by changes in the mean and/or the variability of its properties and that persists for an extended period, typically decades or longer. Climate change may be due to natural internal processes or external forcings such as modulations of the solar cycles, volcanic eruptions and persistent anthropogenic changes in the composition of the atmosphere or in land use.
IPCC Special Report on Global Warming of 1.5 *°C, 2018, Glossary, p.545-546.*

D'après la définition du GIEC (reportée ci-dessus), le changement climatique est un changement d'état du système. N'ayant, en général, pas directement accès aux états, il est identifié par ses effets persistants au vu de l'échelle temporelle de la variable étudiée (température, précipitations, etc.). Il est à noter que cette définition diffère selon les organismes. Par exemple, celle de la Convention-Cadre des Nations Unies sur les Changements Climatiques (CCNUCC) demande, de plus, que le changement identifié soit attribué aux forçages anthropiques. Nous ne ferons pas, dans ce manuscrit, cette distinction car notre but premier est de décrire l'évolution chronologique d'une série d'observations.

1.4 Évolution de la compréhension du climat

Plusieurs éclairages sont ou ont été utilisés pour mieux appréhender les propriétés du climat. Il sont généralement très inter-dépendants et ne représentent pas un changement de paradigme en soi. Nous présenterons ici quelques axes qui ne constituent pas une classification exhaustive

des sciences climatiques, mais plutôt des façons d'aborder la science présentant des liens étroits avec les sujets étudiés durant cette thèse.

1.4.1 Classification climatique

Le point de vue, peut être le plus ancien, est d'associer à différentes zones géographiques du globe, un climat. Cette approche a connu un âge d'or avec les travaux de Köppen [83]. Le but de cette perspective étant d'étudier et, dans le cas de Köppen, de classifier les climats en fonction des régions géographiques comme le montrent les figures 1.5 et 1.6. Cette vision des choses est cependant autrement plus compliquée à mettre en place dans un climat non-stationnaire. En effet, les régions représentatives de certains climats changent. Cet axe reste néanmoins un moyen d'analyse intéressant du climat. Köppen fournit une classification basée sur des indices de températures moyennes à proximité du sol, au pas de temps journalier, dans le but d'expliquer la distribution de la faune et de la flore à la surface du globe [83]. Les indices sont basés sur des périodes caractéristiques/prédéfinies durant lesquelles les températures restent dans un intervalle donné (durant une année). Pour la figure 1.5, les seuils sont 10 °C et 20 °C et les périodes caractéristiques sont 1 et 4 mois.

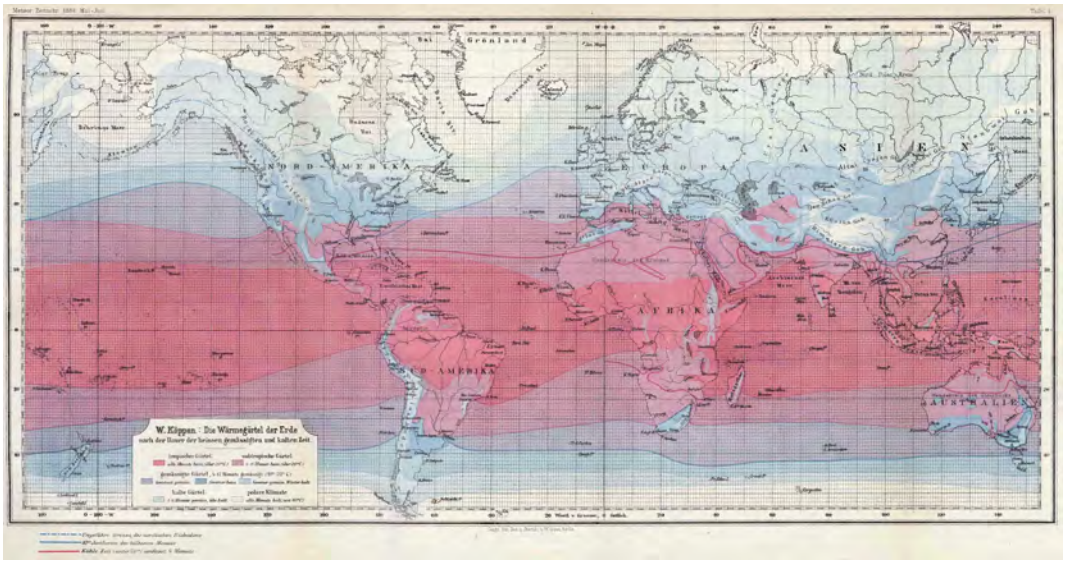


FIGURE 1.5 – Classification due à Köppen 1884.

Ce premier graphique décrit le climat en fonction d'une seule variable, ceci est insuffisant pour caractériser les climats. En effet, les différentes régions sont très peu dépendantes de

leurs longitudes, ce qui semble irréaliste. Geiger [45] propose alors de pousser l'idée un peu plus loin en rajoutant une variable du cycle hydrologique : la précipitation. Le résultat de cette classification est montré dans la figure 1.6.

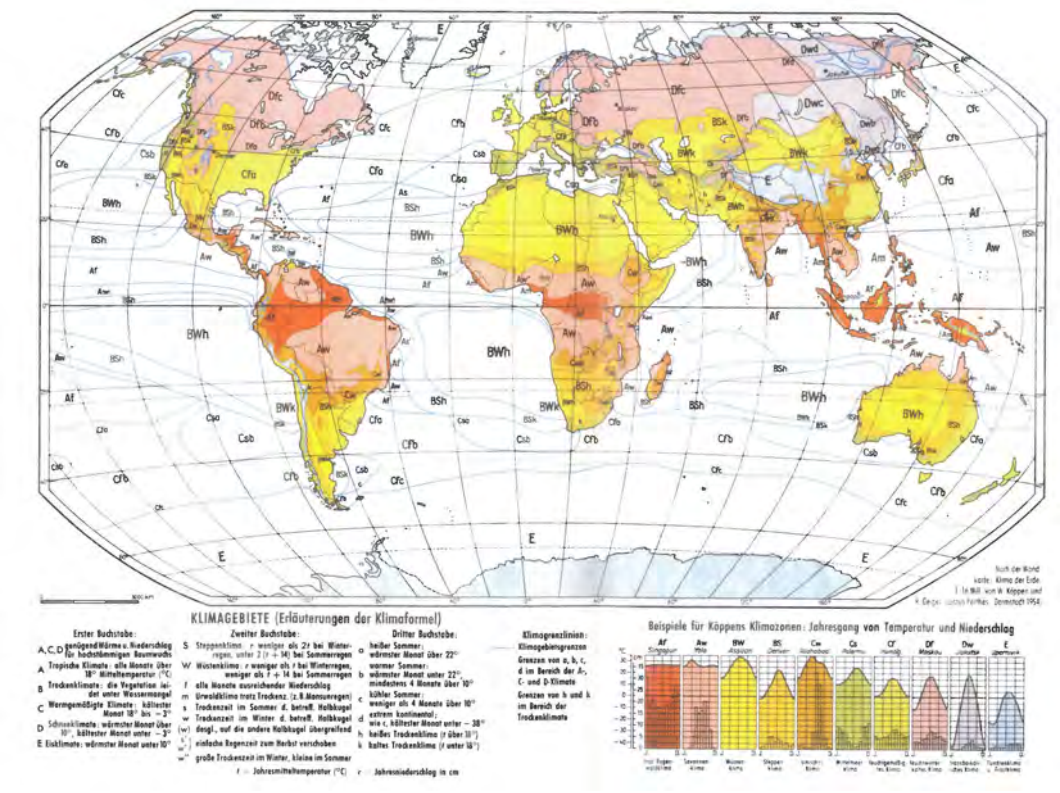


FIGURE 1.6 – Cette carte est une reproduction d'une classification due à Geiger et Köppen datant de 1954. La classification discrimine entre les différents climats à l'aide des cycles annuels de précipitations et de température.

1.4.2 Analogues climatiques

Très proche de ces notions, et toujours dans un thème géographique, viennent les analogues climatiques. Il s'agit de rechercher pour un lieu x_0 au temps t_0 des lieux x_1 présentant un climat similaire au temps t_1. La question est donc de replacer sur le globe le climat (futur ou passé) d'une localisation donnée par le(s) lieu(x) ayant les caractéristiques climatiques les plus proches, et ce pour une période donnée. Par exemple, quelle(s) localisation(s) aurait le cycle saisonnier de températures (en climat actuel) le plus similaire à celui de Toulouse en 2100 ? Cette notion, déjà présente dans la littérature (e.g Kopf *et al.*, 2008), est essentiellement

utilisée pour représenter le changement climatique. En effet, dans leur article de 2008, Kopf *et al.*[82] proposent de replacer à l'aide de ce concept, les climats futurs de villes européennes dans le climat actuel. La figure 1.7 ci-dessous montre les analogues du climat futur de Paris en 2100 calculés à l'aide d'une distance définie sur les distributions de trois variables climatiques : *Degré Jour Unifié (DJU)*, *Degrés Jours de Climatisation (DJC)* et un indice d'aridité due à Thornthwaite [82]). Ces variables mesurent les contraintes climatiques des villes tant en terme énergétique (chauffage et climatisation) qu'en terme de déficit en eau durant les mois arides. En effet, DJU (resp. DJC) est une mesure de l'écart moyen, sur l'année, des températures en-dessous (resp. au-dessus) d'une température de référence fixée à 18 °C. L'indice d'aridité se définit comme la somme sur les mois arides des déficits d'eau relatifs à l'Evapo-Transpiration Potentielle (ETP) i.e (ETP-précipitations)/ETP. Les climats proches de celui de Paris en 2100 sont indiqués par des couleurs chaudes. Le meilleur analogue est atteint en Espagne près de la frontière portugaise dans la ville de Badajoz.

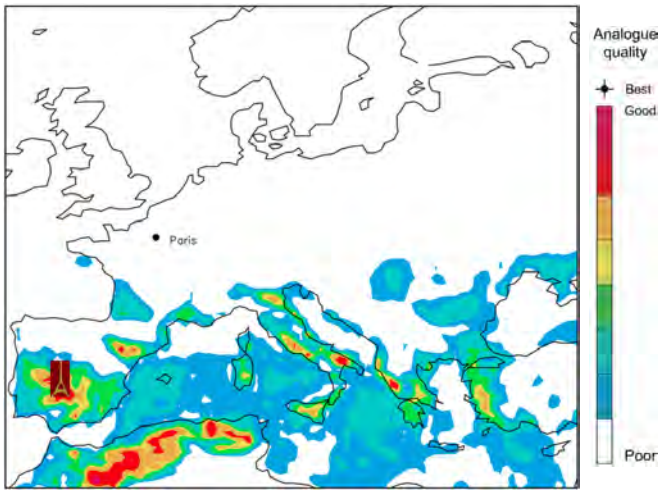


FIGURE 1.7 – Analogues actuelles du climat futur de Paris en 2100 (source Kopf *et al.*, 2012[82]). Plus les couleurs sont chaudes plus les climats sont proches du climat estimé.

1.4.3 Modélisation physique du système climatique

L'approche apportée par la modélisation physique du système climatique est sûrement celle qui a connu le plus de développements, notamment avec l'avènement des modèles de climat. Les modèles de climat sont une version "simplifiée" du système climatique dans le sens où l'on choisit de modéliser les principaux phénomènes du système à l'aide d'équations physiques qui

sont, pour la plupart, discrétisées (figure 1.8). Les objectifs d'une telle démarche sont multiples. Ils permettent aux chercheurs de faire plusieurs expériences *in silico*, généralement sur de grandes échelles de temps et d'espace, permettant d'accroître la compréhension des mécanismes climatiques. Cette méthodologie permet, en outre, de faire des expériences idéalisées, par exemple, en imposant une évolution de variables influant sur le système climatique telle que la concentration de gaz à effet de serre dans l'atmosphère. Le degré de simplification de tels modèles peut être hiérarchisé en fonction de la question posée :

- Energy Balance Models (EBMs) introduit par Budyko (1969) et Sellers (1969). Ces modèles ont été construits pour reproduire le bilan énergétique (radiatif) de la terre. Ils n'ont, en général, pas de description spatiale mais permettent, par exemple, de quantifier la réponse de la température moyenne globale pour un scénario d'accroissement de CO_2.

- Earth-system Models of Intermediate Complexity (EMICs) ou Modèle Terre à Complexité Intermédiaire. À la (grande) différence des EBMs, ces derniers incluent une représentation spatiale (simplifiée) du système "Terre", donc de ses composantes atmosphériques et océaniques au minimum. Leur atout principal est de permettre d'étudier le climat sur des échelles de temps longues, on peut ainsi étudier l'impact de la répartition des continents sur le climat.

- General Circulation Models, Modèles de Circulation Générale (GCMs) ou encore leur évolution : Earth System Models (ESMs). Ils constituent, aujourd'hui, les outils les plus complets pour la réalisation des projections climatiques du $21^{\text{ème}}$ siècle à l'échelle globale. Ceci, tant en terme de précision spatiale, que de représentation de processus dynamiques, physiques et, dans le cas des ESMs, bio-géo-chimiques.

  Initialement, les GCM désignent des modèles ayant trois dimensions spatiales résolvant la dynamique et la physique d'une seule composante du système climatique, par exemple, l'atmosphère (AGCM) ou l'océan (OGCM). Pour résoudre numériquement les équations décrivant l'état de l'atmosphère (ou de l'océan) celui-ci est divisé en un grand nombre de petits volumes dans lesquels les variables sont supposées constantes (ou, en tout cas, admettent une représentation simplifiée). Les phénomènes dynamiques et physiques, principalement issus de la thermodynamique et de la mécanique des fluides, sont alors

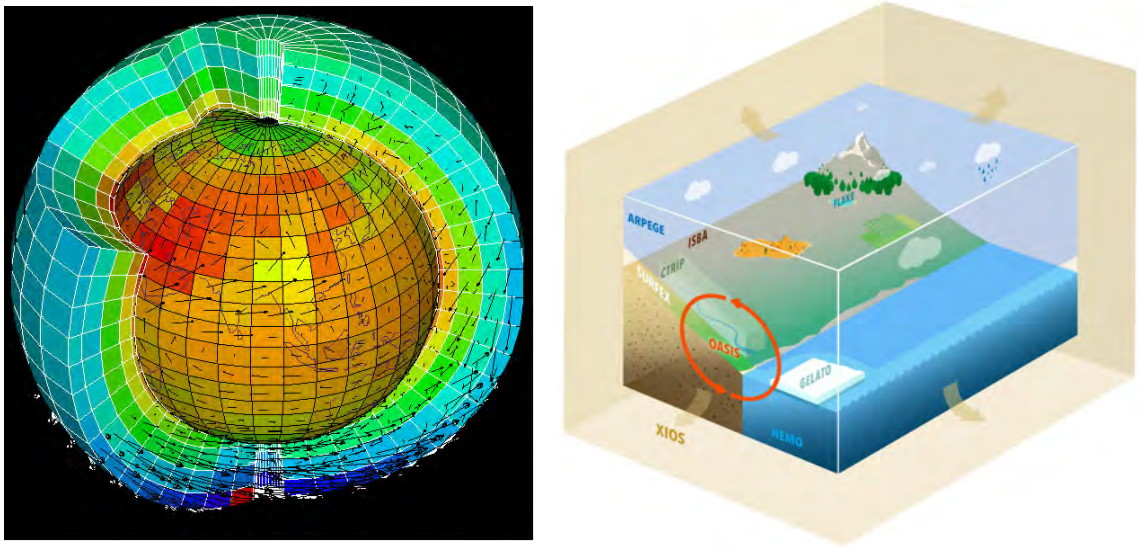


FIGURE 1.8 – Représentation d'un modèle de climat : la figure de gauche montre la discrétisation tridimensionelle (latitude, longitude et composante verticale) faite par un modèle atmosphérique. Cet éclaté représente l'état de l'atmosphère à un instant donné. Le vent est représenté par les flèches noires et la température décroissant du rouge au bleu.
Source : Laboratoire de Météorologie Dynamique (LMD) `http://www.lmd.jussieu.fr/~jldufres/Exposes/IM12-images_v0.pdf`. La figure de droite montre les différentes composantes et interactions représentées par le modèle couplé CNRM-CM6 [126].

approchés par des méthodes de type volumes finis. Cela permet d'obtenir l'évolution des variables sur chacun des volumes de la maille (cf figure 1.8).

Ensuite, les connaissances des flux et interactions entre océan et atmosphère ont permis de mettre en place des modèles résolvant 'simultanément' les deux composantes et leurs principaux échanges. On obtient alors un modèle de circulation générale océan-atmosphère (AOGCM)[59]. Nous appellerons, par abus de langage, *modèle couplé*, tout modèle résolvant simultanément deux composantes. Avec le temps, les modèles couplés ont pris en compte plus de composantes importantes du système (sol, cryosphère ...) représenté dans la figure 1.9, jusqu'à arriver aux "Modèles du Système Terre" (ESM) résolvant de plus, les processus bio-géo-chimiques [14]. Ces derniers permettent de décrire la biosphère ainsi que la chemosphère et donc d'obtenir la répartition d'espèces chimiques dans le système. Ils offrent, en général, une représentation du cycle du carbone.

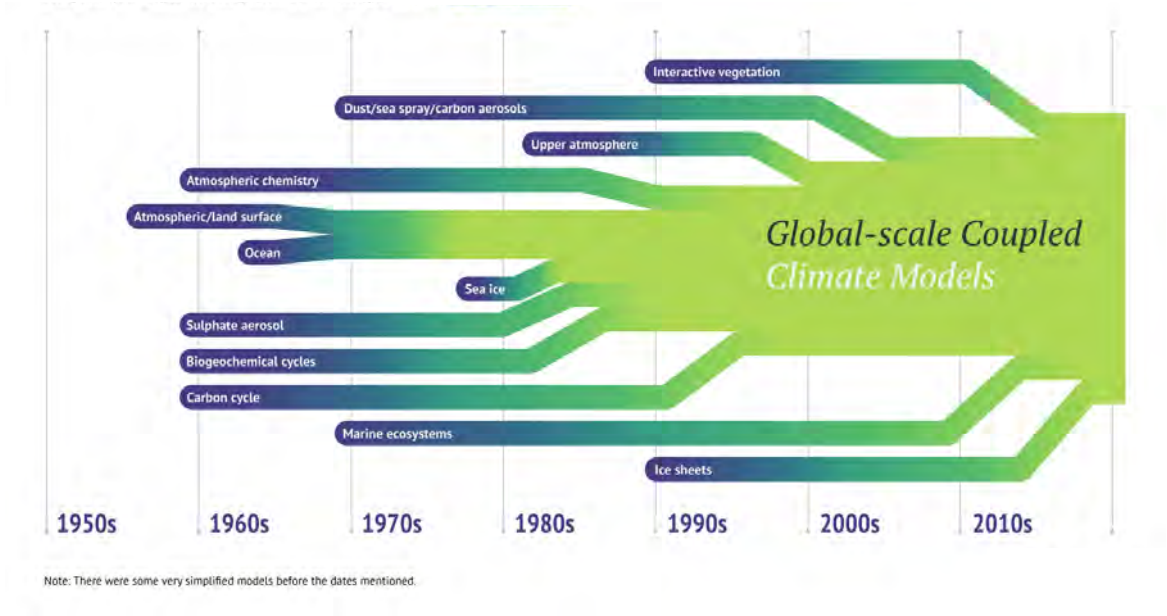


FIGURE 1.9 – Évolution de la complexité des modèles de climat (figure basée sur les travaux du Dr Gavin Schmidt). Cette frise chronologique montre l'évolution des modèles de climat. Ils commencent avec une description de l'atmosphère AGCM puis océan-atmosphère AOGCM pour finalement se coupler progressivement avec plusieurs modélisations d'autres composantes du système climatique.

1.4.4 L'importance des statistiques dans l'étude du climat

Le système climatique est un système dynamique chaotique. L'étude de tels systèmes, qu'ils soient considérés comme déterministes ou purement aléatoires, se fait en général au travers de leurs statistiques [64]. C'est une des raisons pour lesquelles les statistiques du système climatique y jouent un rôle central, le climat pouvant être vu comme l'analyse statistique de ses variables. En effet, on cherche à obtenir différentes informations sur le système climatique ainsi qu'à quantifier leurs incertitudes. Par exemple : y a-t-il un changement dans l'état du système climatique ? (Détection). Est-ce imputable à l'activité humaine ? (Attribution). Avec quelle précision peut-on quantifier l'évolution d'une variable ? C'est dans ce dernier point de vue que nous nous placerons pour étudier des séries chronologiques durant la majorité de ce manuscrit.

Les précédents points développés au cours des sections 1.4.1, 1.4.2, 1.4.3 peuvent être appréhendés dans le cadre des statistiques. En effet, la classification climatique (section 1.4.1)

cherche à répondre à la question : quels sont les grands types de climat ? Ce qui revient à partitionner à l'aide de méthodes d'apprentissage non-supervisé. Le raisonnement avec les analogues de la section 1.4.2 est, quant à lui, à rapprocher d'une estimation par plus proches voisins. Enfin, les relations entre la modélisation physique et les statistiques sont nombreuses et apparaissent à plusieurs niveaux. Une première remarque est que les sorties des ESM visent à reproduire une réalisation du climat. Une des grandes forces des modèles réside dans la possibilité de rejouer plusieurs fois une situation climatique et d'obtenir alors, à l'aide de méthodes statistiques, une meilleure compréhension de la distribution des variables étudiées. En outre, la bonne utilisation des modèles pour analyser le climat est soumise à une évaluation de ces derniers. La validation, la calibration et parfois l'initialisation des modèles couplés donnent lieu à des problèmes d'apprentissage (au sens statistique). Cela revient à dire en termes probabilistes : peut-on reproduire les propriétés statistiques des observations, et comment inférer les paramètres du modèle au travers de l'assimilation de données [106, 19].

1.5 Modélisation de l'évolution saisonnière du climat

Le but premier de ce travail de thèse est d'étudier la déformation des cycles saisonniers de variables climatiques, considérées au pas de temps quotidien, sous l'influence du changement climatique. Nous effectuerons cette analyse sur deux caractéristiques des variables étudiées : l'espérance et les quantiles.

Nous serons donc amenés à estimer, qu'il s'agisse de l'espérance ou des quantiles de la variable, pour une localisation donnée, des fonctions prenant leurs valeurs sur le cylindre. Les cercles représentant l'aspect périodique, portant donc les jours de l'année d, et l'axe principal portant la variable annuelle y. Dans la suite de cette section, nous noterons la fonction d'intérêt

$$\begin{aligned}\Psi\colon \mathbb{S}^1 \times [1, N] &\to \mathbb{R} \\ (d, y) &\mapsto \Psi(d, y)\end{aligned}.$$

Nous supposerons de plus que Ψ satisfait des conditions de régularité qui seront détaillées dans les chapitres 3 et 4.

Un élément intéressant et primordial afin d'étudier le changement climatique saisonnier est de connaître la différence moyenne entre les cycles annuels de deux périodes. Prenons, par

exemple deux périodes de 30 années, débutant aux années y_1 et y_2 ; les périodes s'écrivent alors $P_1 = \{y_1, \ldots, y_1 + 29\}$ et $P_2 = \{y_2, \ldots, y_2 + 29\}$. Autrement dit, si la variable d'intérêt est la moyenne de température $T_{d,y}$ nous modélisons le changement saisonnier comme suit : $T_{d,y_1+t} - T_{d,y_2+t} = h_{_{P_1,P_2}}(d) + \epsilon, \quad t \in [\![0, 29]\!]$. L'élément $h_{_{P_1,P_2}}$ représente alors un cycle annuel donnant le réchauffement expérimenté pour chaque jour de l'année, en moyenne entre les deux périodes, et ε est un bruit centré. Dans la suite, nous utiliserons le terme "delta cycle" pour désigner les cycles portant une information du même type que celle donnée par h. Cette première modélisation ne nous informe cependant pas de l'évolution chronologique annuelle de la déformation du cycle saisonnier. Elle nous montre seulement la différence du comportement moyen entre deux périodes. Ce modèle suppose, de plus, que l'évolution annuelle moyenne à l'intérieur de chaque période est comparable ; une hypothèse qui n'est presque jamais vérifiée pour des périodes opposant un climat stationnaire (e.g. 30 années en climat pré-industriel) à une période subissant un réchauffement (e.g. période actuelle). C'est pour toutes ces raisons qu'il est naturel, lorsque nous nous intéressons à l'intégralité d'une série d'observations sur de longues périodes (60 ans ou plus), de contrôler la magnitude du changement saisonnier par une fonction g dépendant de l'année considérée $g(y)$. Notre modélisation devient alors : $T_{d,y_2} - T_{d,y_1} = (g(y_2) - g(y_1))h(d) + \epsilon$.
L'hypothèse principale faite par ce modèle est que la modulation du changement saisonnier est portée par le cycle h sur l'intégralité de la période. La magnitude du changement est alors gouvernée par la fonction g. Au cours de cette thèse, nous nous référerons à ce modèle comme un modèle multiplicatif ou bilinéaire. Cette approche a déjà porté ses fruits pour la modélisation de signaux temps-espace, il est alors mieux connu sous le nom de "pattern scaling"[96, 121].

Cette hypothèse constitue une étape importante du travail de modélisation présenté dans ce manuscrit. Elle mérite donc de s'attarder d'avantage sur ses motivations d'un point de vue plus théorique.

D'une part, cette hypothèse peut être vue comme une approximation de type développement limité, si la déformation des cycles est contrôlée par une fonction régulière C des années représentant, par exemple, les émissions anthropiques. Il existe alors une fonction Φ telle que la fonction d'intérêt $\Psi(d, y) = \Phi(d, C(y))$ pour tout $d \in [\![1, 365]\!], y \in [\![1, N]\!]$. Il vient :

$\Phi(d, C(y_1)) - \Phi(d, C(y_0)) \simeq (C(y_1) - C(y_0)) \frac{\partial}{\partial y} \Phi(d, C(y_0))$.

Bien sûr, si les variations de C sont trop grandes ou encore que Φ est trop non-linéaire en sa deuxième variable, l'approximation ne tient plus. Cependant, si l'intérêt est d'obtenir le delta cycle moyen, l'emploi de ce modèle reste justifié.

D'autre part, nous pouvons en réalité décrire l'intégralité d'un signal 2D quelconque en généralisant cette approche. En effet, rien ne nous empêche de décomposer le signal de la fonction mesurée en augmentant le nombre de termes multiplicatifs :
$d \in [\![1, 365]\!], y, p \in [\![1, N]\!], \quad N \in \mathbb{N}$
$T_{d,y} = f(d) + g_1(y)h_1(d) + \cdots + g_p(y)h_p(d)$. Lorsque $p = min(N, 365) - 1$, cette décomposition permet d'interpoler l'ensemble des points. Dans cette thèse, nous serons généralement dans le cas $N \leq 365$. Pour le constater, il suffit de localiser les fonctions g, par exemple, $g_y = \mathbb{1}_{y=y_0}$ ou tout autre base de l'espace discret (splines, Fourier, etc.). Comme nous l'avons précisé, nous n'utiliserons cette décomposition qu'au premier ordre. Pour en retenir une information intéressante, nous supposerons, en général, que la fonction f représente un cycle annuel moyen sur la période considérée, g la tendance moyenne annuelle et, pour des raisons d'identifiabilité du modèle, h sera de moyenne 1. Ceci de sorte à ce que le signal restant $g_2(y)h_2(d) + \cdots + g_p(y)h_p(d)$ soit composé de fonctions dont la moyenne, à un jour fixé d ou à une année y fixée, est nulle sur la période. Les fonctions g_i et h_i, par définition non-corrélées aux termes précédents, peuvent alors être incluses dans le modèle selon leur importance, notamment grâce à une mesure des variations de g_i.

1.6 Quelles données pour évaluer le changement climatique ?

Pour inférer les évolutions du climat, plusieurs types de données sont utilisés par la communauté climatique. Les deux grandes classes sont les observations (*in situ*, radar, satellite, ...) et les modèles se basant sur une modélisation physique de l'évolution du système. Ces derniers sont évalués au vu des performances à reproduire le climat passé. Il existe divers produits combinant les deux points de vue notamment à l'aide d'un modèle contraint par les observations disponibles. C'est, par exemple, le cas des données PIOMAS [143] qui sont une ré-analyse d'aires et de volumes de banquise arctique.

1.6.1 Modèles couplés

Au cours de cette thèse, nous utiliserons des simulations provenant de modèles couplés, plus précisément les bases de données issues du Projet d'Intercomparaison des Modèles Couplés/Coupled Model Intercomparison Project CMIP 5. Ces données contiennent notamment des simulations du climat passé ainsi que du climat futur en fonction de divers scénarios d'émission RCP (Representative Concentration Pathway) et cela pour plusieurs modèles climatiques. Le CMIP est un projet du programme mondial de la recherche sur le climat (WCRP). Ce projet vise à réaliser des simulations climatiques de façon coordonnée entre les différents laboratoires de recherche en climat, permettant ainsi une meilleure estimation et compréhension des différences entre les modèles climatiques [118]. L'exercice permet d'évaluer le degré de réalisme des modèles sur le passé récent, de produire des projections sur le futur proche (2035) et plus lointain (fin du $21^{\text{ème}}$ siècle et au delà). En outre, il a pour but de quantifier l'effet des rétroactions les plus importantes telles que celles impliquées dans le cycle du carbone. Les résultats basés sur ces simulations jouent un rôle majeur dans l'évaluation de l'état des connaissances sur le climat par le GIEC. Nous utiliserons, au cours de cette thèse, la cinquième phase de ce projet. Il implique 20 groupes de climat à travers le monde et comprend plusieurs types de simulations longue échéance, ces dernières sont consignées dans la figure 1.10.

Nous utiliserons principalement les simulations historiques et celles issues des scénarios RCP. Ces scénarios sont souvent décrits en terme d'évolution du forçage radiatif. Ce dernier décrit l'impact d'une perturbation du système sur le bilan radiatif terrestre par rapport à un état de référence (dans notre cas 1750 c.f chapitre 8 [117]). Le bilan radiatif est défini comme la différence entre l'énergie radiative reçue (dont le plus grand contributeur est le soleil) et l'énergie radiative émise par le système climatique. Autrement dit, le forçage radiatif positif représente le différentiel d'énergie du système climatique. Lorsque le forçage radiatif est positif le climat tend à se réchauffer et inversement. Les scénarios RCP, au nombre de 4, sont des trajectoires d'évolution du forçage radiatif jusqu'à l'horizon 2300. Ils sont représentatifs de scénarios de concentration de gaz à effet de serre dans l'atmosphère. La figure 1.11 montre les scénarios RCP, le scénario le plus optimiste suppose une augmentation de la concentration de gaz à effet de serre jusqu'en 2010-2020, date après laquelle les émissions diminuent substan-

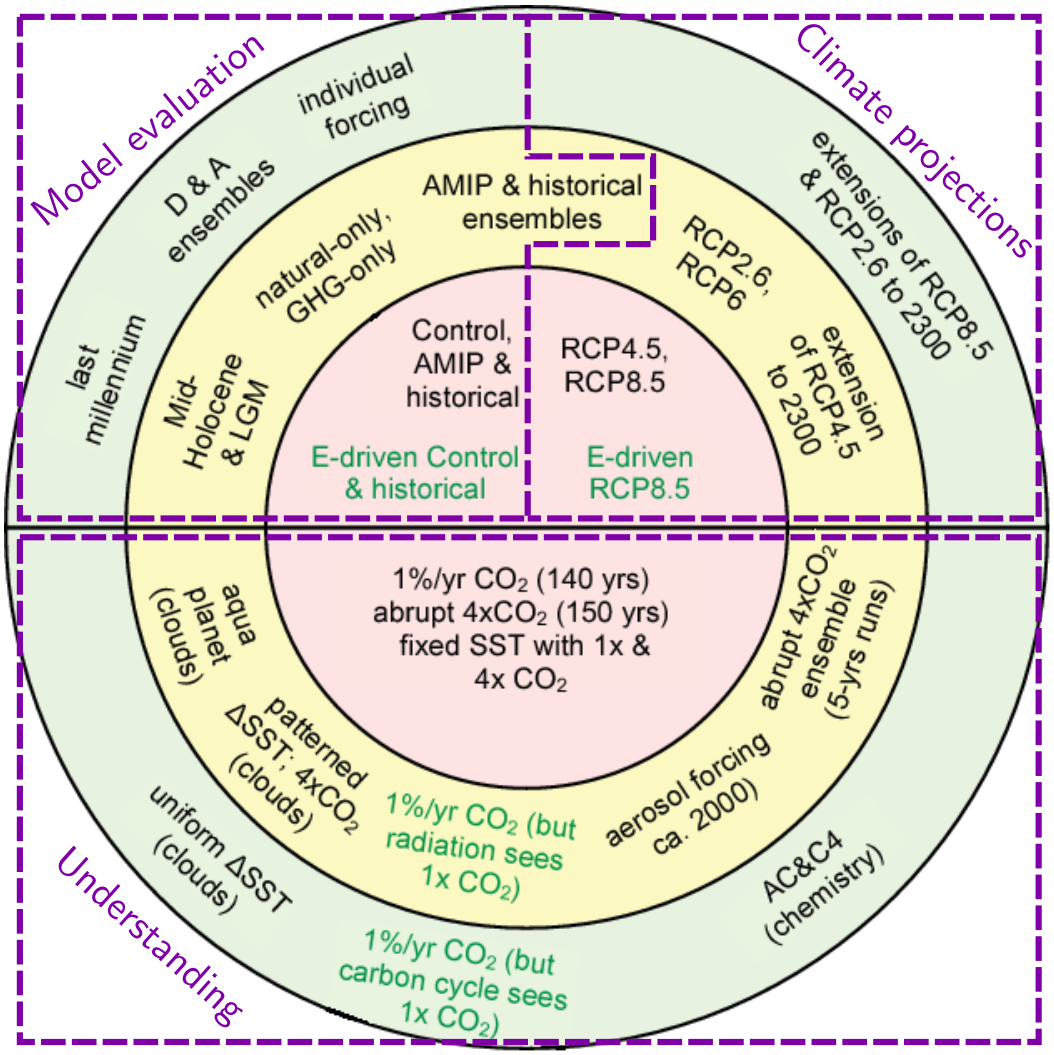


FIGURE 1.10 – Figure adaptée de [119] montrant les différents types d'expériences à long terme du projet CMIP5. Ces expériences peuvent être classifiées en fonction de leurs degrés d'importance au vue du projet : les expériences au coeur du projet (contenue dans le disque rose) et celles du premier et deuxième tiers consignées dans les couronnes jaunes et vertes. Chacune des expériences permet de remplir les objectifs du projet CMIP5. En bas, sont consignées des expériences idéalisées faites pour une meilleure compréhension du climat. Par exemple, l'effet d'une augmentation annuelle de 1% du CO_2 dans l'atmosphère ou encore d'un quadruplement du CO_2. En haut à droite, les simulations basées sur des scénarios d'émission RCP. En haut à gauche, les expériences permettant d'évaluer la qualité des modèles. En vert sont celles effectuées avec des modèles possédant un couplage avec leurs cycles du carbone.

tiellement [122]. De même, pour les scénarios RCP 4.5 et 6 ils atteignent un pic d'émission en 2040 et 2080 puis les émissions déclinent. Le scénario le plus pessimiste RCP 8.5 suppose une augmentation des émissions sur l'intégralité du 21$^{\text{ème}}$.

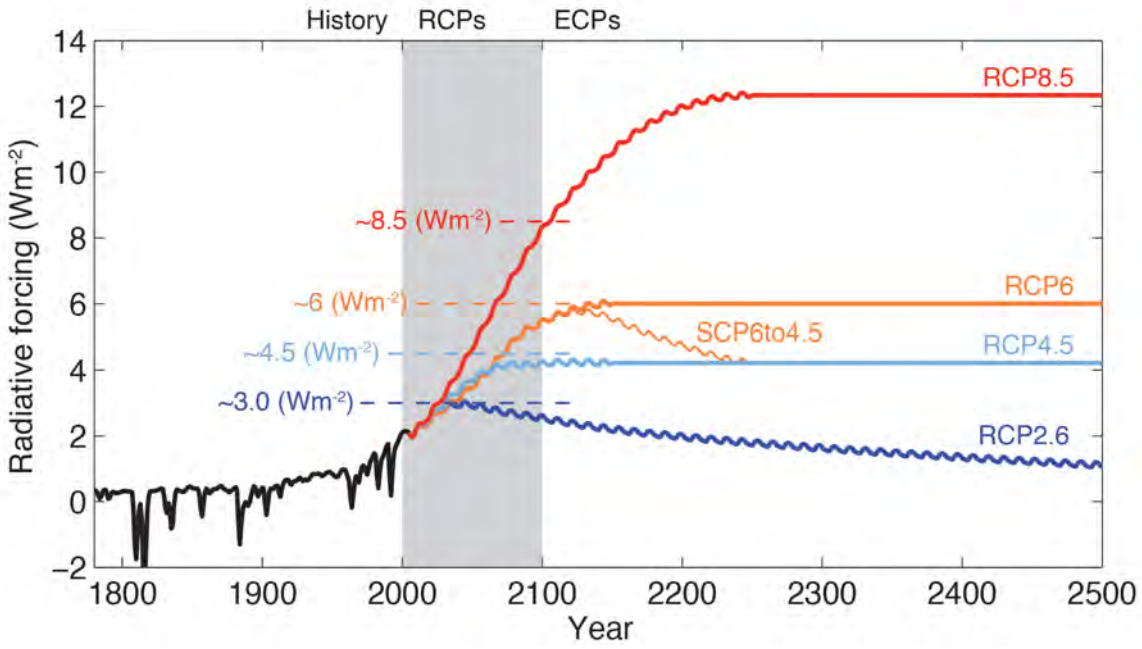


FIGURE 1.11 – Forçages radiatifs imposés par les scénarios d'émission RCP 2.6, 4.5, 6, 8.5 et leurs extensions. Les variations décennales sont principalement dues aux forçages volcaniques et à la prise en compte des cycles solaires. Source : 5ème rapport du GIEC [117].

1.6.2 Les données d'observation

Nous pouvons aussi étudier le changement climatique directement à l'aide d'observations. Il existe un grand nombre de méthodes d'observation du climat à l'aide de différents instruments (cf figure 1.12) : satellites, ballon sonde, balise argo, lidar, avions. Les données *in situ* ayant en général la plus grande profondeur temporelle, nous nous baserons particulièrement sur celles issues de stations météorologiques au sol. Ces dernières ne peuvent, hélas, être utilisées à l'état brut sans avoir subi une analyse de leur qualité et éventuellement des corrections. En effet, les conditions d'observation peuvent changer au cours du temps et causer des ruptures ou autres types de biais dans les séries temporelles observées. En d'autres termes, nous n'avons pas une mesure homogène d'une variable dans le temps à une localisation donnée. Les raisons de ces in-homogénéités peuvent provenir de plusieurs sources non-directement liées au changement climatique, notamment un changement d'instrumentation de mesure ou encore une modification de l'environnement du site de mesure. La figure 1.13 montre l'effet d'un changement d'abri sur des mesures de température journalière. Pour pallier à ce type de désagrément, les

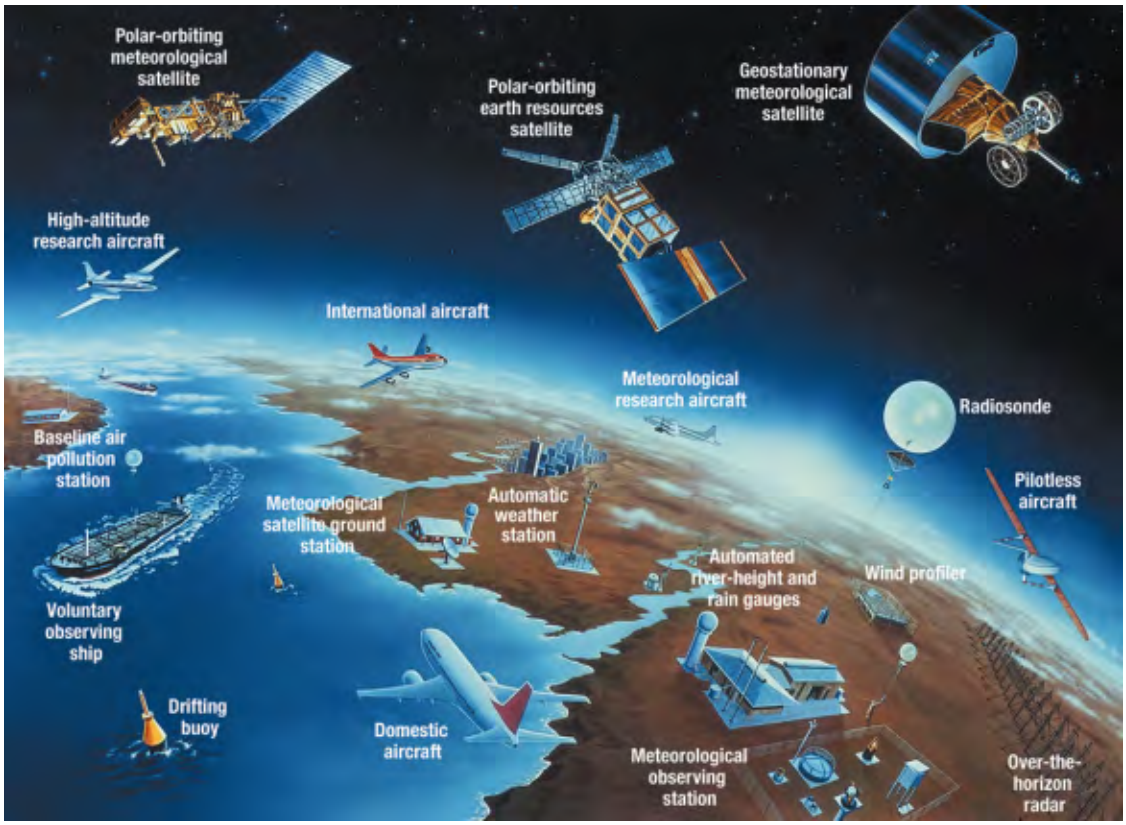


FIGURE 1.12 – Représentation simplifiée de différentes composantes du système mondial intégré de surveillance (WIGOS) de l'OMM [figure extraite du site de Organisation Météorologique Mondial `https://public.wmo.int/en/about-us/vision-and-mission/wmo-integrated-global-observing-system`]

services climatiques utilisent des méthodes de correction des séries (homogénéisation) ou de sélection de stations ne présentant pas ce type de biais.

Nous utiliserons au cours de cette thèse les Séries Quotidiennes de Référence (SQR) qui sont spécifiques à la France. Il s'agit d'une sélection de données climatologiques quotidiennes d'une station météorologique, pour une période donnée. Cette sélection est basée sur diverses informations telle l'amplitude des ruptures mentionnées plus tôt, mais aussi le taux de données manquantes et le nombre de déplacements du poste de mesure.

Il existe un monde reliant ces deux points de vue (modèles et observations) tels que les ré-analyses qui permettent de simuler - avec un modèle physique - le climat d'une période, en contraignant la simulation par les observations disponibles (en général d'un champ assez régulier spatialement : pression, température). Notre objectif étant tout d'abord méthodologique,

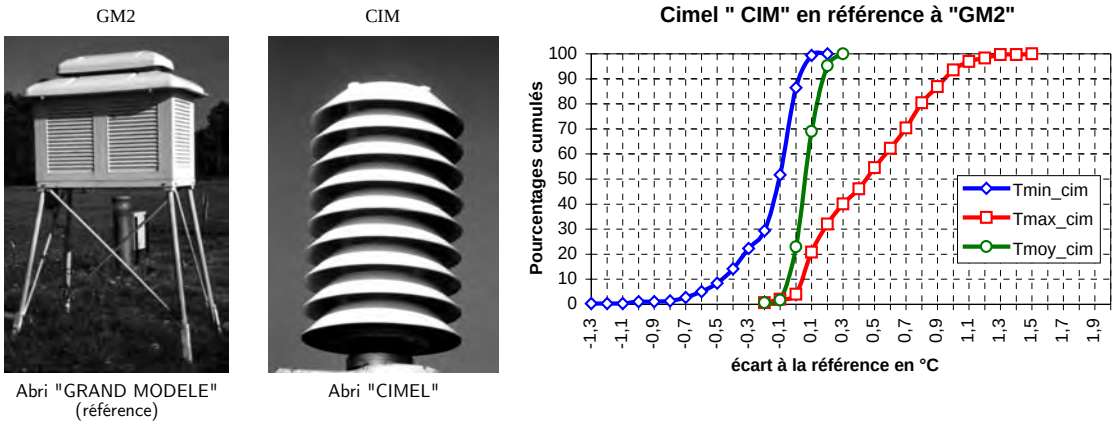


FIGURE 1.13 – Influence de l'abri sur la mesure des températures moyennes minimales et maximales journalières :
Les deux figures de gauche présentent deux types d'abris équipés du même modèle de capteur de température (100 ohms, platine, de classe A). Tous deux positionnés sur le site de Trappes à 11 mètres de distance l'un de l'autre [90]. Il est ici pris comme référence les mesures réalisées en utilisant l'abri « grand modèle ». La figure de droite montre les courbes des fréquences cumulées (exprimées en pourcentage) des écarts à la référence. L'abri « CIMEL », trop petit, a tendance à s'échauffer, ce qui se traduit par un biais d'environ +0.5°C par rapport à l'abri « grand modèle » (médiane des écarts de TX de 0.45°C). En revanche, les températures minimales sont généralement sous-estimées. La température moyenne n'est que peu affectée. Lefèvre (1998) montre l'influence du rayonnement et des classes de vent sur les écarts observés. theSource : thèse d'Olivier Mestre, 2000.

nous ne pousserons pas plus avant le bestiaire des données climatiques.

Ce manuscrit est structuré comme suit. Dans le Chapitre 1, nous proposons une description des théories et outils statistiques utilisées, soit principalement les splines de lissage (au travers de la théorie des RKHS), la régression quantile et quelques aspects de la sélection de modèle. Le Chapitre 2 contient un article publié sur les normales climatiques non-stationnaires ainsi que quelques compléments. Le Chapitre 3 est un article en préparation qui applique, entre autre, des méthodes de régression quantiles permettant de voir l'évolution, sous l'effet du changement climatique, des distributions de la température moyenne et des précipitations sur des observations. Le Chapitre 4 conclut ce travail et présente des perspectives, notamment aux chapitres 2 et 3.

Chapitre 2

Méthodes statistiques

Sommaire

Ce chapitre a pour vocation de présenter, dans un premier temps, un panel choisi de méthodes classiques d'approximations fonctionnelles, permettant ainsi d'introduire les splines de lissage dans un cadre un peu plus large. Ces techniques peuvent s'appliquer à l'estimation de nombreuses caractéristiques d'une variable telles que les quantiles, mais il en sera surtout tiré partie pour estimer l'espérance d'une variable. Dans la deuxième section, sera présentée une introduction à la régression quantile, puis, dans la dernière section, quelques éléments de sélection de modèles pouvant servir à la fois à la bonne estimation du paramètre de lissage (e.g. pour les splines), qu'à la sélection de variables utilisées en régression quantile. Chacune de ces composantes va interagir au cours de ce travail de thèse. Il serait erroné, en revanche, de les penser comme des résumés exhaustifs de ces axes de recherches ; mais plutôt comme un ensemble de résultats à disposition pour mener à bien cette thèse. Pour résumer, nous utiliserons la première et la troisième partie pour obtenir de nouvelles normales offrant plus de flexibilité, et principalement les deux dernières parties pour dériver des distributions en climat changeant. Le peu de théorie des RKHS utilisée dans le chapitre 4 concerne les modèles paramétriques utilisés comme référence. La sélection de la bonne complexité des modèles est au coeur de la plupart des résultats de cette thèse. En effet, elle nous permettra d'obtenir des signaux significatifs, ce qui n'est pas le cas en imposant la complexité du modèle à priori.

2.1 Espaces de Hilbert à noyaux reproduisant (RKHS)

2.1.1 Introduction

Les espaces de Hilbert à noyaux reproduisant (RKHS) prennent une place particulièrement importante en statistique car ils fournissent un cadre général dans lequel on peut offrir la meilleure approximation fonctionnelle (au vu de la norme et des points d'observations). Les RKHS permettent, en outre, de faire des ponts entre différents estimateurs tels que les splines, le krigeage ou encore les séparateurs à vaste marge [42]. Cette théorie a été développée simultanément par Nachman Aronszajn et Stefan Bergman, en 1950 [112], pour estimer la meilleure approximation fonctionnelle de la fonction de régression. Dans cette théorie, on se place dans un espace de Hilbert fonctionnel, généralement de dimension infinie, tel que les fonctionnelles linéaires d'évaluation ($f \mapsto f(x)$) soient continues. Initialement, et, dans le

cadre de cette thèse, cette théorie est motivée par des problèmes de régression pénalisée. La théorie suivante se généralise aisément au cas complexe (noyaux de Bergman). Cependant, par soucis de concision et de clarté, nous ne développerons pas cette partie de la théorie.

2.1.2 Un coup d'œil sur la théorie générale

Définition 2.1 (RKHS)

Soit $\mathbf{X}$ *un ensemble arbitraire et* $\mathcal{H} \subset \mathbb{R}^{\mathbf{X}}$ *un espace de Hilbert de fonctions à valeurs réelles sur* $\mathbf{X}$*. On dit que* $\mathcal{H}$ *est un espace de Hilbert à noyau reproduisant si pour tout* x *dans* $\mathbf{X}$*, les formes linéaires* $L_x \colon \mathcal{H} \to \mathbb{R}$ *sont continues*

$$f \mapsto f(x)$$

i.e. $\forall x \in \mathbf{X},\ \exists M_x$ *tel que* $\forall f \in \mathcal{H}, \|f(x)\| \leq M_x \|f\|_{\mathcal{H}}$

Exemple 2.1

Un espace de Hilbert muni de la norme $\mathcal{L}^2([0,1])$ *contenant les indicatrices* $\sqrt{n}\mathbb{1}_{[0,\frac{1}{n}]}$ *ou encore* $\sqrt{(n - \frac{n^2}{2}x)}\mathbb{1}_{[0,\frac{2}{n}]}$ *(pour une version continue) n'est pas un RKHS. En effet, la norme* L_2 *de ces deux suites de fonctions est constante, et pourtant leurs évaluations en zéro tendent vers l'infini.*

Remarque 2.1

Dans un RKHS, la convergence en norme implique la convergence simple (ou ponctuelle).

Définition 2.2 (Noyau)

Soit $\mathbf{X}$ *un ensemble arbitraire et* $\mathcal{H} \subset \mathbb{R}^{\mathbf{X}}$ *un espace de Hilbert (de fonctions à valeurs réelles sur* $\mathbf{X}$*). On dit que* $\boldsymbol{\mathcal{K}} \colon \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ *est un noyau symétrique positif si*

1. $\forall \{x, y\} \subset \mathbf{X},\ \boldsymbol{\mathcal{K}}(x, y) = \boldsymbol{\mathcal{K}}(y, x)$ *(symétrie).*
2. $\forall n \in \mathbb{N},\ \forall x_1, \ldots, x_n \in \mathbf{X},\ \forall t_1, \ldots, t_n \in \mathbb{R} : \sum_{i,j \in [\![1,n]\!]} t_i t_j \boldsymbol{\mathcal{K}}(x_i, x_j) \geq 0$ *(On dit qu'il est défini lorsque cette dernière inégalité est stricte).*

Remarque 2.2

Hormis la bi-linéarité, il s'agit presque d'un produit scalaire.

Théorème 2.1 (Théorème de Moore-Aronszajn)

La donnée d'un noyau symétrique positif $\boldsymbol{\mathcal{K}}$ est équivalente à celle d'un RKHS sur $\mathcal{H}$.

Preuve:

Une preuve détaillée peut être trouvée dans [48]. Nous nous contenterons, dans ce travail, de donner les idées principales.

$\boxed{\Leftarrow}$ Soit $\mathcal{H}$ muni de son produit scalaire $\langle .,. \rangle_{\mathcal{H}}$. Les formes linéaires L_x étant continues pour tout $x \in \mathbf{X}$, il vient, d'après le théorème de représentation de Riesz :
$\exists K_x \in \mathcal{H}$, tel que $L_x(.) = \langle K_x, . \rangle_{\mathcal{H}}$. On définit alors :

$$\begin{aligned} \boldsymbol{\mathcal{K}} \colon \mathbf{X} \times \mathbf{X} &\to \mathbb{R} \\ (x, y) &\mapsto \langle K_x, K_y \rangle_{\mathcal{H}}. \end{aligned}$$

On vérifie sans peine que $\boldsymbol{\mathcal{K}}$ est un noyau symétrique positif. Notons qu'il n'est pas bilinéaire.

$\boxed{\Rightarrow}$ Réciproquement, étant donné un noyau positif $\boldsymbol{\mathcal{K}}$ construisons le RKHS associé :

(a) **Ébauche du RKHS**

Posons $\mathcal{H}_0 \overset{\text{def}}{=} \{\sum_{i=1}^{n} a_i \boldsymbol{\mathcal{K}}(x_i, .) \mid n \in \mathbb{N}, x_1, \ldots, x_n \in \mathbf{X}, a_1, \ldots, a_n \in \mathbb{R}\}$ (i.e. les combinaisons linéaires des représentants d'évaluation).

$\mathcal{H}_0$ peut alors être muni d'un produit scalaire en étendant, par bi-linéarité, les relations
$\langle K_x, K_y \rangle_{\mathcal{H}_0} \overset{\text{def}}{=} \boldsymbol{\mathcal{K}}(x, y)$.

En effet :

- $\langle .,. \rangle_{\mathcal{H}_0}$ est bien défini :
  $\langle f, g \rangle_{\mathcal{H}_0}$ ne dépend pas de la représentation de f et g. Si $g = \sum_{i=1}^{n} a_i \boldsymbol{\mathcal{K}}(x_i, .)$ alors $\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} a_i f(x_i)$.
- $\langle .,. \rangle_{\mathcal{H}_0}$ est bi-linéaire, symétrique et défini positif :
  soit $f = \sum_{i=1}^{n} a_i \boldsymbol{\mathcal{K}}(x_i, .), n \in \mathbb{N}$.
  Alors $\|f\| = 0 \Rightarrow \forall t \in \mathbf{X},\ f(t) = 0$ (Cauchy–Schwarz).

- $\boldsymbol{\mathcal{K}}(x_i, .)$ est un représentant de l'évaluation.

(b) **Complétion de $\mathcal{H}_0$**

La complétion de $\mathcal{H}_0$ est standard, et les représentants de l'évaluation sont toujours donnés par $\boldsymbol{\mathcal{K}}$.

■

Propriété 2.1 (Projection sur un RKHS)
Soient $(\mathcal{H}_0, \langle ., . \rangle_{\mathcal{H}_0})$ un espace de Hilbert de fonctions définies sur $\mathbf{X}$, et $(\mathcal{H}, \langle ., . \rangle_{\mathcal{H}_0})$ un sous-espace fermé. Alors, si $\mathcal{H}$ est un RKHS de noyau $\boldsymbol{\mathcal{K}}$ la projection de $g \in \mathcal{H}_0$ sur $\mathcal{H}$ est donnée par l'application :

$$\begin{aligned} g_K : \mathbf{X} &\to \mathbb{R} \\ x &\mapsto \langle g, \boldsymbol{\mathcal{K}}(x, .) \rangle_{\mathcal{H}_0}. \end{aligned}$$

Preuve:
Notons $g_{\mathcal{H}}$ le projeté de g sur $\mathcal{H}$.
$\forall x \in \mathbf{X}$,

$$\langle g - g_{\mathcal{H}}, \boldsymbol{\mathcal{K}}(x, .) \rangle_{\mathcal{H}_0} = \langle g, \boldsymbol{\mathcal{K}}(x, .) \rangle_{\mathcal{H}_0} - \langle g_{\mathcal{H}}, \boldsymbol{\mathcal{K}}(x, .) \rangle_{\mathcal{H}_0} = g_K(x) - g_{\mathcal{H}}(x).$$

Or, les $\boldsymbol{\mathcal{K}}$(x,.) sont des éléments de $\mathcal{H}$ et, par définition de la projection orthogonale, il vient :
$\forall x \in \mathbf{X}, \quad g_K(x) - g_{\mathcal{H}}(x) = 0$ ■

Propriété 2.2 (Décompositions en somme de RKHS [110])
Soit $\mathcal{H}$ un espace auto-reproduisant sur $\mathbf{X}$ dont le noyau $\boldsymbol{\mathcal{K}}$ peut se décomposer en la somme de deux noyaux $\boldsymbol{\mathcal{K}} = \boldsymbol{\mathcal{K}}_0 + \boldsymbol{\mathcal{K}}_1$, tel que : $\boldsymbol{\mathcal{K}}_0(x, .) \in \mathcal{H}$ et $\boldsymbol{\mathcal{K}}_1(x, .) \in \mathcal{H}$ pour tout $x \in \mathbf{X}$ et $\langle \boldsymbol{\mathcal{K}}_0(x, .), \boldsymbol{\mathcal{K}}_1(y, .) \rangle = 0$, pour tout $x, y \in \mathbf{X}$. Alors $\mathcal{H}$ est la somme orthogonale $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$ (où $\mathcal{H}_0$ et $\mathcal{H}_1$ sont les RKHS associés aux noyaux $\boldsymbol{\mathcal{K}}_0$ et $\boldsymbol{\mathcal{K}}_1$).
Réciproquement, si $\mathcal{H}_0 \cap \mathcal{H}_1 = 0$, alors $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$ a pour noyau reproduisant $\boldsymbol{\mathcal{K}} = \boldsymbol{\mathcal{K}}_0 + \boldsymbol{\mathcal{K}}_1$.

Preuve:

Par définition de $\mathcal{H}$ et par orthogonalité de $\boldsymbol{\mathcal{K}}_0$ et $\boldsymbol{\mathcal{K}}_1$, il vient :

pour tout $x, y \in \mathbf{X}$, $\quad \boldsymbol{\mathcal{K}}_0(x,y) = \langle \boldsymbol{\mathcal{K}}_0(x,.), \boldsymbol{\mathcal{K}}(y,.)\rangle = \langle \boldsymbol{\mathcal{K}}_0(x,.), \boldsymbol{\mathcal{K}}_0(y,.)\rangle$.

Par suite, $\mathcal{H}_0$ admet un supplémentaire orthogonal $\mathcal{H}_0^{\perp}$.

Soit $f \in \mathcal{H}$, alors il existe $f_0 \in \mathcal{H}_0$ et $f_0^{\perp} \in \mathcal{H}_0^{\perp}$.

Il vient $f(x) = \langle \boldsymbol{\mathcal{K}}(x,.), f\rangle = \langle \boldsymbol{\mathcal{K}}_0(x,.), f_0\rangle + \langle \boldsymbol{\mathcal{K}}_1(x,.), f_0^{\perp}\rangle = f_0(x) + \langle \boldsymbol{\mathcal{K}}_1(x,.), f_0^{\perp}\rangle$ ce qui montre que $\boldsymbol{\mathcal{K}}_1$ est le noyau reproduisant de $\mathcal{H}_0^{\perp}$.

La réciproque est triviale. ■

Théorème 2.2 (Le théorème du représentant)

Soient $\mathcal{H}$ un RKHS de noyau défini positif $\boldsymbol{\mathcal{K}}$, et $\{x_1, \ldots, x_n\} \subset \mathbf{X}$ un sous-ensemble de cardinal fini n. Soit $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ croissante par rapport à son dernier argument.

Alors, toute solution f au problème (si elle existe) :

$$\min_{f \in \mathcal{H}} \Psi(f(x_1), \ldots, f(x_n), \|f\|_{\mathcal{H}})$$

s'écrit comme combinaison linéaire des représentants de l'évaluation :

$f = \sum_{i=1}^{n} \alpha_i . \boldsymbol{\mathcal{K}}(x_i,.)$ *où* $\forall i \in [\![1,n]\!], \alpha_i \in \mathbb{R}$.

Preuve:

Supposons qu'il existe un minimiseur f de Ψ.

Considérons $\mathbf{E} = \mathrm{Vect}\{\boldsymbol{\mathcal{K}}(x_i,.), i \in [\![1,n]\!]\}$ qui est un sous-espace vectoriel de dimension finie, donc fermé de $\mathcal{H}$. D'après le théorème du supplémentaire orthogonal [111], $\mathcal{H}$ se décompose comme une somme directe de deux espaces vectoriels orthogonaux $\mathcal{H} = \mathbf{E} \bigoplus^{\perp} \mathbf{E}^{\perp}$. Nous noterons $f_E \in E$ la projection orthogonale de f sur $\mathbf{E}$ et $f_{\perp} \in \mathbf{E}^{\perp}$ son supplémentaire orthogonal.

Il vient d'une part :

$$\begin{aligned} \forall i \in [\![1,n]\!], \quad f(x_i) &= \langle f, \boldsymbol{\mathcal{K}}(x_i,.)\rangle_{\mathcal{H}} \\ &= \langle f_E, \boldsymbol{\mathcal{K}}(x_i,.)\rangle_{\mathcal{H}} + \langle f_{\perp}, \boldsymbol{\mathcal{K}}(x_i,.)\rangle_{\mathcal{H}} \\ &= \langle f_E, \boldsymbol{\mathcal{K}}(x_i,.)\rangle_{\mathcal{H}} = f_E(x_i). \end{aligned}$$

D'autre part, $\|f\| = \|f_E\| + \|f_\perp\| \geq \|f_E\|$.

Par suite, Ψ étant croissante par rapport à son dernier argument (contenant la norme de la fonction), $\Psi(f(x_1), \ldots, f(x_n), \|f\|_{\mathcal{H}}) \geq \Psi(f_E(x_1), \ldots, f_E(x_n), \|f_E\|_{\mathcal{H}})$.
Par conséquent, $f_\perp = 0_{\mathcal{H}}$ et $f \in \mathbf{E}$. ■

Remarque 2.3

Habituellement le critère à minimiser est de la forme :
$\Psi(f(x_1), \ldots, f(x_n), \|f\|_{\mathcal{H}}) = \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda J(\|f\|_{\mathcal{H}})$ *où* L *est la fonction coût considérée* $J : \mathbb{R}^+ \to \mathbb{R}^+$ *est une fonction croissante. Le premier terme décrit l'ajustement aux données, le second contraint la "régularité" de f, le contrôle du compromis est fait à l'aide du paramètre* $\lambda \in \mathbb{R}^+$.

2.1.3 Le cas des splines (1D) sur l'intervalle [0,1]

Les splines de lissage cherchent à décrire la fonction de régression dans l'espace de Sobolev $\mathcal{H}^m = \{f : [0,1] : \to \mathbb{R} : \ f^{(m-1)}$ absolument continue, $f^{(m)} \in \mathcal{L}^2([0,1])\}$ en pénalisant la fonctionnelle par la norme de sa dérivée d'ordre $m \in \mathbb{N}$.

Étant donné des observations $\{(y_i, x_i)_{i \in [\![1,n]\!]}\} \subset \mathbb{R}^2$, $\lambda \in \mathbb{R}$, elles sont solution du problème de minimisation suivant :

$$\begin{aligned} \hat{g} &= \underset{s \in \mathcal{H}^m}{\operatorname{argmin}} \left[\underbrace{\sum_{i=1}^{n} \left(y_i - s(x_i)\right)^2}_{\text{terme d'ajustement aux données}} + \lambda \underbrace{\int_0^1 \left(s^{(m)}(x)\right)^2 dx}_{\text{régularité de } f} \right] \\ &= \underset{s \in \mathcal{H}^m}{\operatorname{argmin}} \left[\underbrace{\|Y - s(X)\|_2^2}_{\text{terme d'ajustement aux données}} + \lambda \underbrace{\left\|s^{(m)}\right\|_{L^2}^2}_{\text{régularité de } f} \right] \end{aligned} \tag{2.1}$$

, où $Y = (y_i)_{i \in [\![1,n]\!]}$, $X = (x_i)_{i \in [\![1,n]\!]}$ et $s(X) = (s(x_i))_{i \in [\![1,n]\!]}$.

Le terme λ contrôle le compromis entre l'ajustement aux données et la complexité de la fonction s. En effet, si $\lambda = 0$, tout interpolateur des données $(y_i, x_i)_{i \in [\![1,n]\!]}$, par exemple

un polynôme d'interpolation, est un minimiseur. À l'inverse, lorsque $\lambda \to \infty$, $\widehat{g}$ tend vers l'estimateur de la régression linéaire simple sur une base de polynômes de degrés au plus $m-1$. L'estimation de λ est en général difficile. Elle est effectuée en minimisant certains critères (e.g. BIC,GCV...), ou encore à l'aide de techniques de validation croisée. Ce problème sera l'objet de la section 2.3.

La stratégie, à partir d'ici, consistera à déterminer la structure de RKHS sur $\mathcal{H}^m$ associé au problème (2.1), de sorte à pouvoir exprimer le terme de régularisation comme une fonction croissante d'une norme et ainsi appliquer un raisonnement analogue au théorème du représentant. Pour ce faire, il est habituel [127, 67] de décomposer $\mathcal{H}^m$ en la somme directe de deux espaces plus simples. Ces derniers sont basés sur la formule de Taylor-Laplace avec reste intégral.

$$\forall x \in [0,1],\ f(x) = \sum_{k=0}^{m-1} f^k(0).\frac{x^k}{k!} + \int_0^1 \frac{(x-u)_+^{m-1}}{m-1!} f^{(m)}(u)du \tag{2.2}$$

où $(t)_+ = \max(0,t)$.

L'équation (2.2) nous pousse à étudier deux facettes du problème. Dans un premier temps : l'espace des fonctions dont l'évaluation est nulle en 0 et dont toutes les dérivées d'ordre inférieur à $m-1$ sont, elles aussi, nulles en 0. Soit

$$\mathcal{H}_0^m = \{f \in \mathcal{L}^2([0,1]) | \quad \forall k \in [\![0, m-1]\!], \\ f^{(k)} \text{absolument continue et } f^{(k)}(0) = 0, f^{(m)} \in \mathcal{L}^2([0,1])\}$$

muni du produit scalaire $\langle f, g\rangle_{\mathcal{H}_0^m} \overset{\text{def}}{=} \int_0^1 f^{(m)}(x).g^{(m)}(x)dx = \langle f^{(m)}, g^{(m)}\rangle_{L^2}$.
Dans un second temps : l'espace des fonctions dont la dérivée d'ordre m est nulle ou, de façon équivalente, des polynômes de degrés au plus $m-1$, que nous noterons :

$$\begin{aligned}\mathbf{P}^m &\overset{\text{def}}{=} \{f \in \mathcal{L}^2([0,1]) | \quad \forall k \in [\![0, m-1]\!],\ f^{(k)} \text{ absolument continue},\ f^{(m)} = 0_{\mathcal{H}^m}\} \\ &= \text{Vect}\{x^k, \quad k \in [\![0, m-1]\!]\}\end{aligned}$$

muni du produit scalaire $\langle f,g\rangle_{\mathbf{P}^m} \overset{\text{def}}{=} \sum_{k=0}^{m-1} f^{(m)}(0).g^{(m)}(0)$.

Ceci définit un espace euclidien qui a toutes les propriétés voulues.

Étude de $\mathcal{H}_0^m$

- $(\mathcal{H}_0^m, \langle,\rangle_{\mathcal{H}_0^m})$ est un espace de Hilbert.

- $\mathcal{H}_0^m$ est un RKHS de noyau reproduisant : $\boldsymbol{\mathcal{K}}^0(x,y) = \int_0^1 \frac{(x-u)_+^{m-1}}{m-1!}.\frac{(y-u)_+^{m-1}}{m-1!}du.$
  En effet, posons :

$$G_m\colon [0,1]^2 \to \mathbb{R}$$
$$(x,y) \mapsto \frac{(x-y)_+^{m-1}}{m-1!}.$$

  D'après la formule de Taylor (2.2) :
  $\forall f \in \mathcal{H}_0^m, \quad \forall x \in [0,1],\ f(x) = \int_0^1 \frac{(x-u)_+^{m-1}}{m-1!} f^{(m)}(u)du = \int_0^1 G_m(x,u).f^{(m)}(u)du.$
  Par suite :
  $\forall f \in \mathcal{H}_0^m, \quad \forall x \in [0,1],\ L_x(f) = f(x) = \int_0^1 G_m(x,u).f^{(m)}(u)du \leq \|f\|_{\mathcal{H}_0^m} \|G_m(x,.)\|_{L^2}.$
  Les fonctionnelles d'évaluation étant continues, $\mathcal{H}_0^m$ est un RKHS. De plus, on trouve par identification le représentant de l'évaluation :
  $f(x) = \langle \boldsymbol{\mathcal{K}}^0(x,.), f\rangle_{\mathcal{H}_0^m} = \int_0^1 \frac{d^m \boldsymbol{\mathcal{K}}^0(x,u)}{du^m}.f^{(m)}(u)du = \int_0^1 G_m(x,u).f^{(m)}(u)du.$
  Par identification $\frac{d^m \boldsymbol{\mathcal{K}}^0(x,u)}{du^m} = G_m(x,u)$. Or, la propriété de reproduction implique :
  $\boldsymbol{\mathcal{K}}^0(x,y) = \langle \boldsymbol{\mathcal{K}}^0(x,.), \boldsymbol{\mathcal{K}}^0(y,.)\rangle_{\mathcal{H}_0^m} = \int_0^1 G_m(x,u).G_m(y,u)du$
  qui est un élément de $\mathcal{H}_0^m$.

Étude de P^m

P^m est clairement un espace euclidien ayant toutes les propriétés voulues. Reste à déterminer le noyau associé à P^m.

Pour ce faire, munissons P^m d'une base orthonormale $\{\phi_k\colon x \to \frac{x^k}{k!}, k \in [\![0, m-1]\!]\}$,

alors : $\forall f \in P^m, \quad \forall x \in [0,1], \quad f(x) = \sum_{k=0}^{m-1} \langle f, \phi_k\rangle_{P^m}.\phi_k(x) = \langle f, \sum_{k=0}^{m-1} \phi_k.\phi_k(x)\rangle_{P^m}$

donc $L_x(.) = \boldsymbol{\mathcal{K}}^P(x,.) = \sum_{k=0}^{m-1} \phi_k(.)\phi_k(x)$.

La propriété de reproduction permet de conclure :

$$\begin{aligned}\boldsymbol{\mathcal{K}}^P(x,y) &= \langle \boldsymbol{\mathcal{K}}^P(x,.), \boldsymbol{\mathcal{K}}^P(y,.)\rangle_{P^m} \\ &= \langle \sum_{k=0}^{m-1} \phi_k.\phi_k(x), \sum_{k=0}^{m-1} \phi_k\phi_k(y)\rangle_{P^m} \\ &= \sum_{k=0}^{m-1} \phi_k(x).\phi_k(y) = \sum_{k=0}^{m-1} \frac{x^k}{k!}.\frac{y^k}{k!}.\end{aligned}$$

$\mathcal{H}^m$ est la somme directe des deux espaces P^m et $\mathcal{H}_0^m$

- $\mathcal{H}^m$ muni du produit scalaire $\langle\rangle_{\mathcal{H}^m} \overset{\text{def}}{=} \langle,\rangle_{P^m} + \langle,\rangle_{\mathcal{H}_0^m}$ se décompose en la somme directe orthogonale de P^m et $\mathcal{H}_0^m$.

- D'après la proposition 2.2, $\mathcal{H}^m$ est un RKHS de noyau $\boldsymbol{\mathcal{K}} = \boldsymbol{\mathcal{K}}^P + \boldsymbol{\mathcal{K}}^0$.

Finalement, $(\mathcal{H}^m), \langle,\rangle_{\mathcal{H}^m}$ est muni d'une structure de RKHS dont on a déterminé le noyau $\boldsymbol{\mathcal{K}}(x,y) = \sum_{k=0}^{m-1} \frac{x^k}{k!}.\frac{y^k}{k!} + \int_0^1 G_m(x,u).G_m(y,u)du$.

Remarque 2.4

Il est à noter que d'autres normes topologiquement équivalentes à celle introduite ici peuvent être considérées pour résoudre le même problème ([127] p7 & section 10.2).

Résolution du problème d'optimisation (2.1)

Théorème 2.3 (Kimeldorf & Wahba, 1971)

Le problème d'optimisation (2.1) *admet une unique solution f. Celle-ci s'écrit sous la forme :* $f = \sum_{k=0}^{m-1} \alpha_k\phi_k + \sum_{i=0}^{n} \beta_i \boldsymbol{\mathcal{K}}^0(x_i,)$.

De plus, le minimiseur est une spline naturelle (i.e dont la dérivée seconde s'annule avant la première observation x_1 et après la dernière observation x_n) et peut s'écrire dans sa base associée $\{n_j(.), i \in [\![1,n]\!]\}$ *:*

$f = \sum_{i=1}^{n} \theta_i n_i(.)$ *où* $\theta = (N^\top N + \lambda\Omega_N)^{-1}N^\top\ Y$ *avec* $N = (n_j(x_i))$, $Y = (y_i)$

et $\Omega_N = (\langle n_i(.), n_j(.)\rangle_{\mathcal{H}_0^m}) = \int_0^1 n_i^{(m)}(x).n_j^{(m)}(x)dx$

Preuve:

- **f est un élément d'un sous-espace E de dimension finie de $\mathcal{H}^m$:**

  Posons $\mathbf{E} = \text{Vect}\{\phi_0, \ldots, \phi_m, \boldsymbol{\mathcal{K}}^0(x_1, .), \ldots, \boldsymbol{\mathcal{K}}^0(x_n, .)\}$. Étant fermé, il admet un supplémentaire orthogonal que nous noterons $\mathbf{E}^\perp$.

  $\forall f \in \mathcal{H}^m,\ \exists!\ \alpha_k, \beta_i \in \mathbb{R}, \rho \in \mathbf{E}^\perp$ tel que $f = \sum_{k=0}^{m-1} \alpha_k \phi_k + \sum_{i=0}^{n} \beta_i \boldsymbol{\mathcal{K}}^0(x_i, .) + \rho$.

  Or, par construction, $\forall i \in [\![1, n]\!], \langle \boldsymbol{\mathcal{K}}(x_i, .), \rho \rangle_{\mathcal{H}^m} = 0$ (car $\boldsymbol{\mathcal{K}}(x_i, .) \in \mathbf{E}$)
  et $\|f\|^2_{\mathcal{H}_0^m} = \left\|\Pi_{\mathcal{H}_0^m}(f)\right\|^2_{\mathcal{H}^m} = \left\|\sum_{i=0}^{n} \beta_i \boldsymbol{\mathcal{K}}^0(x_i, .)\right\|^2_{\mathcal{H}^m} + \|\rho\|^2_{\mathcal{H}^m}$ (par Pythagore).

  La fonction à minimiser se ré-écrit donc comme suit :
  $\Psi(f) = \sum_{i=1}^{n} (y_i - \sum_{k=0}^{m-1} \alpha_k \phi_k(x_i) + \sum_{j=0}^{n} \beta_i \boldsymbol{\mathcal{K}}^0(x_j, x_i))^2 + \left\|\sum_{i=0}^{n} \beta_i \boldsymbol{\mathcal{K}}^0(x_i, .)\right\|^2_{\mathcal{H}^m} + \|\rho\|^2_{\mathcal{H}^m}$.
  Par conséquent un minimiseur de Ψ est un élément de $\mathbf{E}$.

- **Écriture sous forme matricielle**

  Posons $Y = (y_i)_{i \in [\![1,n]\!]}$, $\Phi = (\phi_k)_{k \in [\![0,m-1]\!]}$, $K = (\boldsymbol{\mathcal{K}}^0(x_i, x_j))_{i,j \in [\![1,n]\!]}$. Alors, le problème (2.1) revient à minimiser :
  $\Psi(\alpha, \beta) = \|Y - \Phi\alpha - K\beta\|_2^2 + \lambda \beta^\top \mathbf{K} \beta$ où $\alpha \in \mathbb{R}^m, \beta \in \mathbb{R}^n$.

  Différencions Φ par rapport à α et β :

$$\begin{aligned} &\nabla_\alpha \Psi = \Phi^\top (Y - \Phi\alpha - K\beta) = 0 \\ \Rightarrow \quad &\Phi^\top \Phi \alpha = \Phi^\top (Y - K\beta) && (2.3) \\ \Rightarrow \quad &\alpha = (\Phi^\top \Phi)^{-1} \Phi^\top (Y - K\beta). && (2.4) \end{aligned}$$

  On obtient ainsi la régression linéaire par moindres carrés par Φ sur Y privé de sa composante en K : $(Y - K\beta)$.

$$\begin{aligned} &\nabla_\beta \Psi = -K^\top (Y - \Phi\alpha - K\beta) + \lambda K\beta = 0 \\ \Rightarrow \quad &(K + \lambda I_n)\beta = (Y - \Phi\alpha) && (2.5) \\ \Rightarrow \quad &\beta = (K + \lambda I_n)^{-1} (Y - \Phi\alpha). && (2.6) \end{aligned}$$

  (Estimateur ridge généralisé avec produit scalaire $x^\top K x$ sur Y privé de sa composante en Φ.)

En multipliant (2.5) par $\Phi^\top$, il vient : $\Phi^\top((K+\lambda I_n)\beta+\Phi\alpha) = \Phi^\top Y$. Puis, en injectant cette équation dans (2.3) il s'ensuit :

(2.3) $\Phi^\top\Phi\alpha = \Phi^\top Y - \Phi^\top K\beta = \Phi^\top((K+\lambda I_n)\beta+\Phi\alpha) - \Phi^\top K\beta$

$\Rightarrow \quad \Phi^\top\beta = 0.$

Finalement, en multipliant (2.6) par $\Phi^\top$:

$\Phi^\top(K+\lambda I_n)^{-1}Y = \Phi^\top(K+\lambda I_n)^{-1}\Phi\alpha$

$\Rightarrow \quad \alpha = (\Phi^\top(K+\lambda I_n)^{-1}\Phi)^{-1}\Phi^\top(K+\lambda I_n)^{-1}Y.$

Puis, en substituant α dans l'équation (2.6) on obtient β :

$$\begin{aligned}\beta &= (K+\lambda I_n)^{-1}(Y - \Phi(\Phi^\top(K+\lambda I_n)^{-1}\Phi)^{-1}\Phi^\top(K+\lambda I_n)^{-1}Y)\\ &= (K+\lambda I_n)^{-1}(In - \Phi(\Phi^\top(K+\lambda I_n)^{-1}\Phi)^{-1}\Phi^\top(K+\lambda I_n)^{-1})Y.\end{aligned}$$

- **f est une spline naturelle**

  - Il est clair que sur les intervalles $[x_i, x_{i+1}]_{i\in[\![1,n-1]\!]}$, f est un polynôme de degré au plus $2m-1$ (étant l'intégrale d'un polynôme de degré $2m-2$).
  - f est de degré au plus $m-1$ sur $[x_n, 1]$ car pour tout $t > x_n$ et pour tout $i \in [\![1,n]\!]$ $\boldsymbol{\mathcal{K}}^0(x_i,t) = \int_0^1 \frac{(x_i-u)_+^{m-1}}{m-1!}.\frac{(t-u)_+^{m-1}}{m-1!}du = 0.$
  - f est de degré au plus $m-1$ sur $[0, x_1]$. En effet, $\Phi^\top\beta = 0$. Alors $\forall t \le x_1 \quad \sum_{i=1}^n \beta_i.\boldsymbol{\mathcal{K}}^0(x_i,t) = \int_0^1 \frac{(t-u)^{m-1}}{m-1!}.\sum_{i=1}^n \beta_i.\frac{(x_i-u)^{m-1}}{m-1!}du = 0$ car $\sum_{i=1}^n x_i^k\beta_i = 0$ pour tout $k \in [\![0, m-1]\!]$.

- **Ré-écriture dans la base spline naturelle**

  $\Psi(\theta) = \|Y - N\theta\|_2^2 + \lambda\theta^\top\Omega_N\theta$ où $\theta \in \mathbb{R}^n$.

  Ψ est strictement convexe et admet donc un unique minimum qui peut être obtenu en annulant la différentielle de Ψ :

$$\begin{aligned}\nabla\Psi(\theta) &= -N^\top(Y - N\theta) + \lambda\Omega_N\theta = 0\\ \Rightarrow \theta &= (N^\top N + \lambda\Omega_N)^{-1}N^\top Y.\end{aligned}$$

■

Propriété 2.3 (matrice de lissage S_λ)
Notons $S_\lambda \overset{\text{def}}{=} N(N^\top N + \lambda\Omega_N)^{-1}N^\top$ *la matrice de lissage spline. Alors :*

- S_λ *est symétrique définie positive.*
- $S_\lambda = (I_n + \lambda K)^{-1}$*, où* K *ne dépend pas de* λ *(forme de Reinsch [42]).*
- $\exists U \in \mathcal{O}_n$ *tel que* $S_\lambda = UDU^\top$*, où* $D = \text{diag}(\frac{1}{1+\lambda d_k}, k \in [\![1, n]\!])$*, les* d_k *étant les valeurs propres de la matrice* K*.*
- $S_\lambda^2 \preceq S_\lambda$ *(i.e.* $\forall x \in \mathbb{R}^n \quad x^\top S_\lambda^2 x \leq x^\top S_\lambda x$*).*
- *Le problème* (2.1) *peut être re-paramétré à nouveau :* $\text{argmin}_\theta\{\|Y - U\theta\|_2^2 + \lambda\theta^\top D\theta\}$*.*
- *Les deux premières valeurs propres sont égales à 1.*

2.1.4 RKHS et base hilbertienne

Un sous-ensemble conséquent de problèmes de la forme (2.1), peuvent être traités de façon simple lorsqu'une base hilbertienne est à disposition et lorsque nous sommes dans les hypothèses d'un théorème spectral - tel que celui dû à Mercer pour les opérateurs compacts, présenté ci-dessous. On supposera dans cette section $\mathcal{L}^2(\mathbf{X})$ séparable (i.e. admettant une base dénombrable).

Théorème 2.4 (Mercer-Hilbert-Schmidt[108, 21])
Soient $\boldsymbol{\mathcal{K}} \in \mathcal{L}^2(\mathbf{X}^2)$ *un noyau continu. À* $\boldsymbol{\mathcal{K}}$*, on associe l'opérateur intégral* T_K *défini par :* $[T_K\varphi](x) = \int_{\mathbf{X}} K(x, s)\varphi(s)\,\mathrm{d}s$*.*
Alors, il existe une base hilbertienne constituée des fonctions propres $(\phi_i)_{i\in\mathbb{N}} \in \mathcal{L}^2(\mathbf{X})$ *de* T_K *de valeurs propres associées* $(\lambda_i)_{i\in\mathbb{N}}$ *positives :*

$$[T_K(\phi_i)](s) = \langle \boldsymbol{\mathcal{K}}(s, .), \phi_i\rangle_{\mathcal{L}^2(\mathbf{X})} = \lambda_i\phi_i(s)$$
$$\boldsymbol{\mathcal{K}}(s, t) = \sum_{i=1}^{\infty} \lambda_i\phi_i(s)\phi_i(t)$$
$$\|\boldsymbol{\mathcal{K}}\|^2_{\mathcal{L}^2(\mathbf{X}^2)} = \sum_{i=1}^{\infty} \lambda_i^2 < \infty.$$

Corollaire 2.1

Sous les hypothèses du théorème 2.4, le RKHS associé à $\boldsymbol{\mathcal{K}}$ *est donné par :*
$\mathcal{H}_{\mathcal{K}} = \{f \in \mathcal{L}^2(\mathbf{X}) | \, \|f\|_{\mathcal{K}}^2 \leq \infty\}$ *où* $\|f\|_{\mathcal{K}}^2 = \sum_{i=1}^{\infty} \frac{\langle f, \phi_i \rangle_{\mathcal{L}^2(\mathbf{X})}^2}{\lambda_i}$.

Preuve:

$\mathcal{H}_{\mathcal{K}}$ est clairement un espace de Hilbert. Il reste à montrer que $\boldsymbol{\mathcal{K}}$ est bien son noyau reproduisant :

$$\begin{aligned} \forall x \in \mathbf{X}, \quad \forall f \in \mathcal{H}_{\mathcal{K}}, \quad \langle \boldsymbol{\mathcal{K}}(x,), f \rangle_{\mathcal{K}} &= \sum_{i=1}^{\infty} \frac{\langle f, \phi_i \rangle_{\mathcal{L}^2(\mathbf{X})} . \langle K(x,), \phi_i \rangle_{\mathcal{L}^2(\mathbf{X})}}{\lambda_i} \\ &= \sum_{i=1}^{\infty} \frac{\langle f, \phi_i \rangle_{\mathcal{L}^2(\mathbf{X})} . \lambda_i . \phi_i(x)}{\lambda_i} \\ &= \sum_{i=1}^{\infty} \langle f, \phi_i \rangle_{\mathcal{L}^2(\mathbf{X})} . \phi_i(x) = f(x). \end{aligned}$$

$\forall x \in \mathbf{X}, \quad \boldsymbol{\mathcal{K}}(x, .) \in \mathcal{H}_{\mathcal{K}}$ car

$$\begin{aligned} \|K(x,)\|_{\boldsymbol{\mathcal{K}}} &= \sum_{i=1}^{\infty} \frac{(\lambda_i . \phi_i(x))^2}{\lambda_i} \\ &= \sum_{i=1}^{\infty} \lambda_i . \phi_i(x)^2 = K(x, x) \leq \infty. \end{aligned}$$

Enfin, les fonctionnelles d'évaluation sont continues :

$$|f(x)| = |\langle \boldsymbol{\mathcal{K}}(x, .), f \rangle_{\boldsymbol{\mathcal{K}}}| \leq \|\boldsymbol{\mathcal{K}}(x, .)\|_{\boldsymbol{\mathcal{K}}} \|f\|_{\boldsymbol{\mathcal{K}}} .$$

■

Le théorème suivant donne une classe de problèmes de régularisation dont la solution s'exprime comme une somme finie de fonctions n'influant pas sur le critère de régularité, et une combinaison linéaire des représentants de l'évaluation.

Théorème 2.5 (d'existence des solutions [49])

Soit $\mathcal{H}$ *un espace de Hilbert sur les fonctions de* $\mathbb{R}^d$ *à valeurs dans* $\mathbb{R}$. *Le minimiseur du*

problème

$$\min_{f\in\mathcal{H}}\Bigg[\underbrace{\sum_{i=1}^{n}L(y_i,f(x_i))}_{\text{terme d'ajustement aux données}}+\lambda\underbrace{J(f)}_{\text{régularité de } f}\Bigg]$$

où $J(f)=\int_{\mathbf{R}^d}\frac{|\mathcal{F}(f)(x)|^2}{\mathcal{F}(g)(x)}dx$ *avec* $\mathcal{F}(g)(x)\geq 0$ *et* $\mathcal{F}(g)(x)\xrightarrow[\|x\|\to\infty]{}0$, $\mathcal{F}(g)$ *est symétrique,* $\ker(J)=\{f\in\mathcal{H}|J(f)=0\}=\mathrm{Vect}(\phi_k,\quad k\in[\![1,K]\!])$ *est de dimension finie, L est une fonction coût,* $\lambda\in\mathbb{R}^+$ *paramètre permettant le compromis entre ajustement aux données et régularité de f,* $(x_1,y_1),\ldots,(x_N,y_N)\in\mathbb{R}^{d+1}$*. s'écrit sous la forme :*

$$f(x)=\sum_{k=1}^{K}\alpha_k\phi_k(x)+\sum_{i=1}^{N}\beta_i g(x-x_i)$$

où $\forall k\in[\![1,K]\!]\quad\alpha_k\in\mathbb{R},\quad\forall i\in[\![1,N]\!]\quad\beta_i\in\mathbb{R}$

De plus si $L(x,y)=(x-y)^2$*, les vecteurs* $\alpha=(\alpha_k),\quad\beta=(\beta_i)$ *vérifient :*

$$\Phi\alpha+(G+\lambda I_N)\beta=Y$$
$$\Phi\beta=0$$

où $(G)_{ij}=g(x_i-x_j)$, $(\Phi)_{ik}=\phi_k(x_i)$.

Remarque 2.5

Définir J de la sorte peut sembler assez restrictif. Cependant, J pénalise les composantes de haute fréquence du signal, ce qui est en règle générale, une propriété attendue de la fonction à approximer.

Le premier terme, combinaison linéaire de l'espace nul, étant dans la vaste majorité des cas défini comme le zéro d'un opérateur différentiel, est appelé composante polynomiale.

2.1.5 Le cas des splines périodiques (1D) sur l'intervalle [0,1]

Le cas périodique peut se traiter de manière analogue au cas d'un intervalle de la droite réelle. En effet, on considère :

$$\mathcal{H}^m(per)=\{f\colon[0,1]:\to\mathbb{R}:\forall k\in[\![0,m-1]\!],\ f^{(k)}\text{absolument continue et périodique,}$$

$f^{(m)} \in \mathcal{L}^2([0,1])\}$ munie du la norme $\|f\|^2_{\mathcal{H}^m(per)} = (\int_0^1 f(u)du)^2 + \int_0^1 (f^{(m)}(u))^2 du$.
Notons que f est considérée périodique lorsqu'elle se raccorde en 0 et en 1. On décompose $\mathcal{H}^m(per)$ en la somme directe de deux RKHS : $\mathcal{H}^m(per) = \{1\} \bigoplus \mathcal{H}_0^m(per)$ où

- $\{1\}$ désigne les fonctions constantes munies de la norme $\|f\|_c^2 = (\int_0^1 f(u)du)^2$.

- $\mathcal{H}_0^m(per)$ le sous-espace de $\mathcal{H}^m(per)$ composé des fonctions d'intégrale nulle munies de la norme $\|f\|^2_{\mathcal{H}_0^m(per)} = \int_0^1 (f^{(m)}(u))^2 du$.

Le noyau reproduisant de $\{1\}$ est la fonction constante égale à 1.
Celui de $\mathcal{H}_0^m(per)$ peut être décrit à l'aide des polynômes de Bernoulli $B_m([31])$, ces derniers sont définis par récurrence :
$\forall t \in [0,1], \quad B_0(t) = 0, \quad \forall r \in \mathbb{N}^* \quad \frac{1}{r} B_r'(t) = B_{r-1}(t), \quad \int_0^1 B_r(t)dt = 0.$
Le noyau s'écrit alors : $\boldsymbol{\mathcal{K}}^0(x,y) = \dfrac{(-1)^{m-1}}{2m!} B_{2m}(\lfloor x-y \rfloor)$
où $\lfloor x-y \rfloor$ désigne la partie factionnaire de $x-y$. Puis, en utilisant la même méthodologie, on obtient que le minimiseur est une spline périodique (polynôme par morceaux).
Pour obtenir cette expression explicite du noyau de $\mathcal{H}_0^m(per)$, il est instructif de le décrire dans la base hilbertienne usuelle de $\mathcal{H}_0^m(per)$, c'est-à-dire, la base de Fourier.
En effet, $\forall f \in \mathcal{H}_0^m(per), \quad t \in [0,1[, \quad f(t) = \sqrt{2}\sum_{k=0}^{\infty} a_k \cos(2\pi kt) + b_k \sin(2\pi kt)$ et
$\|f\|^2_{\mathcal{H}_0^m(per)} = \int_0^1 (f^{(m)}(u))^2 du = \sum_{k=0}^{\infty} (2\pi k)^{2m}(a_k^2 + b_k^2)$.

$$\langle \boldsymbol{\mathcal{K}}(x,.), f\rangle_{\mathcal{H}_0^m(per)} = \sum_{k=1}^{\infty} (2\pi k)^{2m}(\alpha_k(x)a_k + \beta_k(x)b_k) = \sqrt{2}\sum_{k=1}^{\infty} a_k \cos(2\pi kx) + b_k \sin(2\pi kx).$$

Il vient : $\alpha_k(x) = \frac{\sqrt{2}}{(2\pi k)^{2m}} \cos(2\pi kx)$ et $\beta_k(x) = \frac{\sqrt{2}}{(2\pi k)^{2m}} \sin(2\pi kx)$.

Puis, par la propriété de reproduction :

$$\begin{aligned}
\boldsymbol{\mathcal{K}}(x,y) &= \langle \boldsymbol{\mathcal{K}}(x,.), \boldsymbol{\mathcal{K}}(y,.) \rangle_{\mathcal{H}_0^m(per)} \\
&= \sum_{k=1}^{\infty} (2\pi k)^{2m} (\alpha_k(x)\alpha_k(y) + \beta_k(x)\beta_k(y)) \\
&= \sum_{k=1}^{\infty} \frac{2}{(2\pi k)^{2m}} (\cos(2\pi kx)\cos(2\pi ky) + \sin(2\pi kx)\sin(2\pi ky)) \\
&= \sum_{k=1}^{\infty} \frac{2}{(2\pi k)^{2m}} (\cos(2\pi k(x-y))).
\end{aligned}$$

Ce développement en série de Fourier est à rapprocher à celui des polynômes de Bernoulli donné dans [1] p 805, ou encore [6] p 61 : $B_{2m}(x) = (-1)^{m-1} 2(2m)! \sum_{k=1}^{\infty} \frac{\cos(2\pi kx)}{(2\pi k)^{2m}}$ ce qui mène au résultat voulu.

2.2 Régression quantile

Depuis l'article fondateur de Koenker et Bassett [77] la régression quantile est devenue un outil important pour obtenir des informations précises sur la distribution conditionnelle d'une variable d'intérêt. Les estimateurs en résultant permettent, en général, une description plus riche que les modèles de régression classiques. En outre, cette statistique est plus robuste, par exemple, à la présence de valeurs extrêmes, que l'estimateur des moindres carrés. En effet, les changements de densité au-dessus et en-dessous du quantile considéré, n'impactent pas ces estimateurs ; c'est une variable locale - propriété qui n'est pas partagée par les moments. Néanmoins, le considérable essor de ces techniques, durant ces 40 dernières années [76], est principalement dû, d'une part aux progrès effectués dans la résolution des programmes linéaires et, d'autre part aux résultats concernant l'inférence pour ces estimateurs. En effet, historiquement, les premiers travaux sur la médiane peuvent être datés de 1760. Bošković [116] propose d'ajuster une droite minimisant la valeur absolue des résidus, il fournit pour ce faire un algorithme géométrique. Ce problème était motivé par l'estimation de l'ellipticité de la terre et apparaît un demi-siècle avant les publications de Gauss [44] et Legendre [91] sur les moindres carrés. Plus tard, Pierre-Simon Laplace en 1789 [73], reprit ces travaux avec sa "méthode de situation" : mélange d'estimateurs de la moyenne (pour l'intercept) et de la médiane (pour la pente) dans son "Traité de mécanique céleste" et étudie plus en profondeur les estimateurs associés.

Par la suite, la question reste intouchée pendant plus d'un siècle jusqu'à ce que Edgeworth en 1888 [36] propose d'étudier le problème de régression médiane tel qu'on le connaît aujourd'hui et, anticipant les travaux de Tukey[54], montre que la variance asymptotique de l'estimateur peut être plus petite que celle de la moyenne (il considère des données issues de mélanges Gaussien). Cependant l'algorithme proposé ne permet pas de résoudre le problème au-delà de la dimension deux.

Après cela,[73] peu de développements sur le sujet sont à noter jusqu'aux travaux fondateurs de Roger Koenker à la fin des années 70. Dans un premier temps, nous décrirons les théorèmes de base de la régression quantile. Dans un second temps, nous nous concentrerons sur les résultats d'inférence de l'estimateur.

2.2.1 Définition et premières propriétés

À la recherche de la fonction coût

L'objet de cette section est d'obtenir les quantiles d'une variable aléatoire $\mathbf{Y}$ conditionnellement aux variables du vecteur aléatoire $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_k)$ comme le minimiseur d'un problème d'optimisation. Au cours de cette section, nous distinguerons les réalisations des variables aléatoires par deux notations différentes. Lorsqu'une seule réalisation est effectuée, elle sera notée en lettres minuscules ; par exemple y désignera une réalisation de $\mathbf{Y}$, de même pour $x_1, \ldots, x_k$ et pour les variables $\mathbf{X}_1, \ldots, \mathbf{X}_k$. Lorsque plusieurs réalisations sont en jeu, nous noterons $Y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ et $X = (x_{i,j}) \in M_{n,k}$ la matrice dont les colonnes sont constituées des n réalisations des régresseurs ou, de façon équivalente, dont les lignes sont composées des réalisations $X_1 = (x_{1,1}, \ldots, x_{1,k}), \ldots, X_n = (x_{n,1}, \ldots, x_{n,k})$ de $\mathbf{X}$ (*design matrix*).

Définition 2.3

Soit $\mathbf{Y}$ une variable aléatoire de fonction de répartition (FDR) $F_{\mathbf{Y}}(z) = P(\mathbf{Y} \leq z)$. Le quantile τ est défini de la manière suivante :

$$Q_{\mathbf{Y}}(\tau) = F_{\mathbf{Y}}^{-1}(\tau) = \inf_y \{y : F_{\mathbf{Y}}(y) \geq \tau\}$$

où $\tau \in]0, 1[$.

Lemme 2.1

$$Q_{\mathbf{Y}}(\tau) \in \underset{u \in \mathbb{R}}{\operatorname{argmin}} \, \mathbb{E}\left[(\rho_\tau(\mathbf{Y} - u))\right]$$

où

$$\rho_\tau(u) \overset{\text{def}}{=} u(\tau - \mathbb{1}(u < 0)) = u(\tau - 1)\mathbb{1}(u < 0) + u\tau\mathbb{1}(u \geq 0)$$

et $\mathbb{1}(u < 0) \overset{\text{def}}{=} \begin{cases} 1 & \textit{si } u < 0, \\ 0 & \textit{sinon.} \end{cases}$

Preuve:

$\min_u \mathbb{E}\left[(\rho_\tau(\mathbf{Y}-u))\right] = \min_u \left\{(\tau-1)\int_{-\infty}^{u}(y-u)dF_{\mathbf{Y}}(y) + \tau\int_{u}^{\infty}(y-u)dF_{\mathbf{Y}}(y)\right\}$.

Puis, après dérivation par rapport à u, un minimiseur q,

doit vérifier $0 = (1-\tau)\int_{-\infty}^{q} dF_{\mathbf{Y}}(y) - \tau\int_{q}^{\infty} dF_{\mathbf{Y}}(y)$.

Donc, $F_{\mathbf{Y}}(q) = \tau$ et $q = Q_{\mathbf{Y}}(\tau)$ est bien un minimiseur. ■

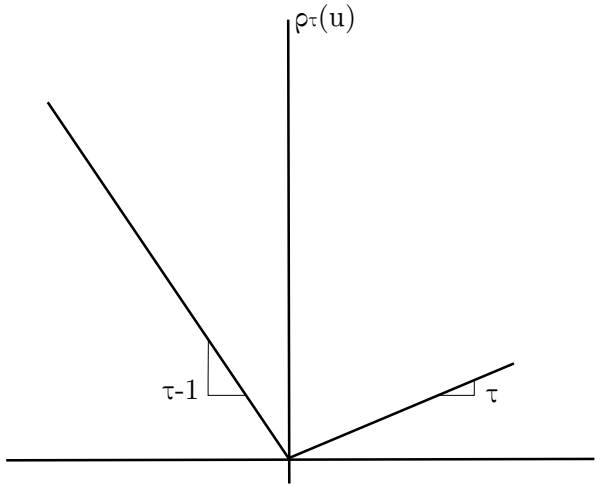


FIGURE 2.1 – La fonction coût associée aux quantiles.

Remarque 2.6

Intuitivement, la fonction coût ρ_τ cherche le point u tel qu'une proportion τ des données soit au-dessus de u.

L'admissibilité de l'estimateur (i.e. il n'existe pas de meilleur estimateur) pour cette fonction coût ρ_τ, a été obtenu dans le cas d'une variable aléatoire admettant un moment d'ordre 2 et une fonction de répartition continue [76, 41].

Ceci amène naturellement à étudier l'estimateur des quantiles suivant.

Proposition 2.1 ([73] p92)

Soient $y_1, \ldots, y_n \in \mathbb{R}$, n réalisations indépendantes d'une variable aléatoire $\mathbf{Y}$ de FDR continue. Alors,

$\hat{q}_\tau \in \underset{q\in\mathbb{R}}{arg\,min} \sum_{i=1}^{n} \rho_\tau(y_i - q) = \underset{q\in\mathbb{R}}{arg\,min} \left[(\tau-1)\sum_{y_i<q}(q-y_i) + \tau\sum_{y_i\geq q}(y_i-q)\right]$.

est un estimateur consistant de $Q_{\mathbf{Y}}(\tau)$.

Remarque 2.7

Le théorème de Mosteller[128] ou encore, une application du théorème de Glivenko-Cantelli [73, 98](fondamental de la statistique), permet de montrer la convergence (vers une loi normale). En revanche, la consistance fait défaut à cet estimateur dès que la fonction de répartition est discontinue. Notons que si τ est irrationnel, le minimum est unique, sinon n'importe quel choix admissible convient.

On cherche à présent à modéliser les quantiles de la variable $\mathbf{Y}$ conditionnellement aux variables $\mathbf{X_i}$. Bien que l'estimation dans un cadre non-linéaire soit possible [80], nous nous cantonnerons, dans un premier temps, à décrire les quantiles par une fonction linéaire en les variables $\mathbf{X_i}$, le modèle s'écrit donc :
$Q_{\mathbf{Y}|\mathbf{X}}(\tau) = \sum_{i=1}^{p} \beta(\tau)_i \mathbf{X_i}, \quad \tau \in]0,1[$.
Ce qui peut être formulé de façon équivalente $Y = \sum_{i=1}^{p} \beta(\tau)_i \mathbf{X_i} + \varepsilon_\tau$ où $Q_{\varepsilon_\tau|\mathbf{X}}(\tau) = 0$.
La régression quantile telle qu'introduite par Koenker and Bassett (1978) [77], propose d'obtenir une estimation de $\beta(\tau)$ comme un minimiseur de la fonction coût ρ_τ :
$\beta(\tau) \in \underset{\beta \in \mathbb{R}^k}{\arg\min} E(\rho_\tau(\mathbf{Y} - \mathbf{X}\beta))$.

L'estimateur des paramètres de la régression quantile $\hat{\beta}_\tau$ est alors obtenu en minimisant l'équivalent empirique :

$$\hat{\beta_\tau} \in \underset{\beta \in \mathbb{R}^k}{\arg\min} \sum_{i=1}^{n} \rho_\tau(y_i - X_i\beta) = \underset{\beta \in \mathbb{R}^k}{\arg\min} \Big[(\tau - 1) \sum_{y_i < X_i\beta} (y_i - X_i\beta) + \tau \sum_{y_i \geq X_i\beta} (y_i - X_i\beta) \Big]. \tag{2.7}$$

2.2.2 Premières propriétés

Équivariances de l'estimateur

Dans cette sous-section, on s'intéresse à la robustesse de l'estimation à la déformation du jeu de données, tout d'abord par déformations linéaires, puis par redistribution par une fonction croissante. Durant tout cette sous-partie, nous noterons $\hat{\beta}_\tau$:
$\hat{\beta}(\tau; Y, X) \in \underset{\beta \in \mathbb{R}^k}{\arg\min} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i\beta)$ et supposerons que l'ensemble des solutions est non-vide.

Équivariance d'échelle

Pour tout $a > 0$ et $\tau \in]0,1[$

$\hat{\beta}(\tau; aY, X) = a\hat{\beta}(\tau; Y, X)$,

$\hat{\beta}(\tau; -aY, X) = -a\hat{\beta}(1-\tau; Y, X)$.

Preuve:

Si $\hat{\beta}$ minimise $\sum_{i=1}^{n} \rho_\tau(Y_i - X_i\beta)$ alors $\hat{\beta}$ minimise $\sum_{i=1}^{n}(\rho_\tau(aY_i - X_i a\beta))$.

Puis, en utilisant le fait que $-a$ change le signe de $(Y_i - X_i\beta)$, le deuxième résultat s'en suit :

$$\begin{aligned}\hat{\beta}(\tau; -aY, X) = & -a.\arg\min_{\beta\in\mathbb{R}^k}\Big[(\tau-1)\sum_{-a.y_i<-a.X_i\beta} -a.(y_i - X_i\beta) \\ & + \tau \sum_{-a.y_i\geq -a.X_i\beta} -a.(y_i - X_i\beta)\Big] \\ = & -a.\arg\min_{\beta\in\mathbb{R}^k}\left[(\tau-1)\sum_{y_i>X_i\beta}(y_i - X_i\beta) + \tau\sum_{y_i\leq X_i\beta}(y_i - X_i\beta)\right].\end{aligned}$$

■

Équivariance à la translation

Pour tout $\gamma \in R^k$ et $\tau \in]0,1[$: $\hat{\beta}(\tau; Y + X\gamma, X) = \hat{\beta}(\tau; Y, X) + \gamma$.

Équivariance à la re-paramétrisation de la matrice de design

Soient $A \in GL_p(\mathbb{R})$ et $\tau \in]0,1[$ alors $\hat{\beta}(\tau; Y, XA) = A^{-1}\hat{\beta}(\tau; Y, X)$.

Invariance par une transformation monotone

Si h est une fonction croissante de $\mathbb{R}$, nous avons la propriété suivante [61] :

$$h(Q_{\mathbf{Y}|\mathbf{X}}(\tau)) \equiv Q_{h(\mathbf{Y})|\mathbf{X}}(\tau).$$

Remarque 2.8

- *L'invariance par une transformation monotone n'est généralement pas vérifiée pour l'estimateur de la moyenne,par exemple si* $\mathbf{X} \sim \mathcal{U}[0,1]$ *:*

  $\log(0.5) = \log(\mathbb{E}[\mathbf{X}]) \neq \mathbb{E}[\log(\mathbf{X})] = -1.$

- *Cette dernière propriété, qui porte sur les quantiles et non leurs estimateurs, est très*

utilisée dans certains modèles non-linéaires comme les modèles à censures. Par exemple, lorsque l'on se place dans un modèle de régression quantile $Q_{\mathbf{Y}|\mathbf{X}}(\tau) = \mathbf{X}\beta(\tau)$ *mais que l'on a à notre disposition uniquement des variables tronquées* $y_i^* = max(0, y_i)$ *alors, en utilisant l'invariance par transformation monotone, on peut montrer que :* $\hat{\beta_\tau} = \underset{\beta \in \mathbb{R}^k}{arg\ min} \sum_{i=1}^n \rho_\tau(y_i - max(0, X_i\beta))$ *est un estimateur consistant [105] de la variable latente y du modèle. Il est à noter que le modèle est alors non-linéaire et que la complexité des calculs informatiques est rédhibitoire. En pratique, on préfère conduire l'estimation en se ramenant au modèle de régression sur un sous-ensemble d'observations que l'on sait non-censurées [28].*

2.2.3 Estimation : écriture sous la forme d'un problème linéaire

Il n'existe pas de solution explicite de (2.7), il faut donc le résoudre numériquement. De plus, la fonction objectif n'est ni différentiable en tout point, ni strictement convexe. Heureusement, le problème de minimisation (2.7) peut être reformulé dans le cadre de l'optimisation linéaire. En programmation linéaire, il s'agit de minimiser une forme linéaire sous des contraintes, elles aussi, définies à l'aide d'applications linéaires. Le problème s'écrit sous forme matricielle :

$$\begin{cases} \inf_{x \in \mathbb{R}_+^n} c^\top x \\ Ax = b, \end{cases} \quad \text{où } c, x \in \mathbb{R}^n, b \in \mathbb{R}^m, A \in M_{m,n}(\mathbb{R}).$$

(Cette formulation du problème est aussi appelée *forme standard*.)

Pour résoudre ce type de problème, deux grandes familles d'algorithmes existent : l'algorithme du simplexe et celui de points intérieurs. Le premier parcourt les sommets du polyèdre $\mathbf{P} = \{x \in \mathbb{R}^n | Ax \leq b\}$ dans la direction de plus forte pente et le second utilise une méthode de gradient relaxé pour approximer la solution.

Posons $r = (Y - X\beta_\tau) \in \mathbb{R}^n$, $r^+ = (\max(0, r_i))_{i \in [\![1,n]\!]}$, $r^- = \max(0, -r_i)_{i \in [\![1,n]\!]}$ de sorte que $r = r^+ - r^-$. Avec les notations précédentes l'équation (2.7) peut être ré-écrite :

$$(\hat{r^+}, \hat{r^-}, \hat{\beta_\tau}) = \underset{\beta \in \mathbb{R}^k, r^\mp \in \mathbb{R}^{+n}}{\arg\min} \left\{ (1-\tau)\mathbf{1}^\top r^- + \tau \mathbf{1}^\top r^+ \mid r^+ - r^- = (Y - X\beta) \right\}. \tag{2.8}$$

Autrement dit, il faut trouver les vecteurs $r_+, r_- \in \mathbb{R}^{+n}$ et $\beta \in \mathbb{R}^k$

qui minimisent : $\left[\tau\mathbf{1}^\top, \quad (1-\tau)\mathbf{1}^\top, \quad 0\right] \begin{bmatrix} r^+ \\ r^- \\ \beta \end{bmatrix}$

sous la contrainte : $\left[I_n, \quad -I_n, \quad X\right] \begin{bmatrix} r^+ \\ r^- \\ \beta \end{bmatrix} = Y.$

Remarque 2.9

- *Soient $r^+, r^- \in \mathbb{R}^{+n}$ issus d'une solution du problème* (2.8)*. Alors :*

  $\forall i \in [\![1, n]\!]$ $\min(r_i^+, r_i^-) = 0$. *En effet, si* $\min(r_i^+, r_i^-) = \mu \neq 0$ *alors* $r_i^{\mp} - \mu$ *est une meilleure solution.*

- *Une forme duale du problème linéaire est proposée dans l'appendice [77]. Celle-ci permet une meilleure optimisation du programme linéaire, mais aussi de généraliser les rangs.*

Théorème 2.6 (Existence des solutions [77])
Si $rang(X) = k$ alors il existe un sous-ensemble de k observations, $h \subset [\![1, n]\!]$ $card(h) = k$ tel que la solution $\widehat{\beta}$ du problème (2.7) *s'écrit :*
$\widehat{\beta}(\tau) = X(h)^{-1}Y(h)$ *où $X(h)$ et $Y(h)$ sont les sous-matrices de X, Y composées des lignes d'indices h.*
De plus, l'ensemble des solutions est décrit par l'enveloppe convexe de solutions de cette forme.

Preuve:
La preuve provient directement de la forme du problème d'optimisation. Pour plus de détail, voir appendice 1 de [77] ou encore [2]. ■

Remarque 2.10
La régression quantile interpole donc un nombre fini d'observations k, en termes de programmation linéaire, cela correspond à vérifier exactement k contraintes, donc être sur un sommet du polyèdre $\mathbf{P}$.

Dans le théorème précédent, nous avons établi un résultat d'existence des solutions. L'unicité, quant à elle, nécessite des hypothèses sur la régularité des fonctions de répartitions F_{y_i}.

Théorème 2.7 (Unicité [77])

Soient $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ *variables aléatoires admettant pour densités* $f_1, \ldots, f_n$ *et* $\mathbb{E}\left[\sum_{i=1}^n f_i(Q_{\mathbf{Y}}(\tau)).X_i^\top X_i\right] = D_1(\tau)$ *où* D_1 *est symétrique définie positive, alors le problème* (2.8) *admet une unique solution* $\widehat{\beta}_\tau$ *avec probabilité* 1.

Propriété 2.4

Notons respectivement Z, N, P le nombre de composantes nulles, positives et négatives de $Y - X\widehat{\beta}$ *où* $\widehat{\beta}$ *une solution du problème* (2.8)*. Alors* $\frac{N}{n} \leq \tau \leq 1 - \frac{P}{n}$

Remarque 2.11

En d'autre termes, la propriété précédente nous assure que à $\frac{Z}{n}$ *près, une proportion* τ *de données se situe en-dessous du quantile estimé, et* $1 - \tau$ *au-dessus.*

2.2.4 Théorie asymptotique

Les propriétés asymptotiques de $\widehat{\beta}_\tau$ sont délicates à établir et, à notre connaissance, il n'existe pas de résultats généraux sur ce sujet. Cependant, dans de nombreux cas (IID,INID,...) on obtient la convergence en loi de l'estimateur $\widehat{\beta}_\tau$ vers une loi normale en imposant, d'une part des conditions de régularité sur les densités au voisinage du quantile considéré et, d'autres part des conditions (plus classiques) sur les co-variables ([50] p12 & [73] p95). Dans la suite, nous nous placerons dans le cadre de données indépendantes mais non nécessairement identiquement distribuées (INID).

Consistance de l'estimateur [73, 38]

Dans leur article de 1999, Bantli & Hallin [38], apportent des conditions nécessaires et suffisantes à la consistance de l'estimateur $\hat{\beta}(\tau)$:

1. $\forall \varepsilon \in \mathbb{R}_+, \quad \sqrt{n}(a_n(\varepsilon) - \tau) \xrightarrow[n\to\infty]{} \infty, \quad \sqrt{n}(b_n(\varepsilon) - \tau) \xrightarrow[n\to\infty]{} \infty$
   où $a_n(\varepsilon) = \frac{1}{n}\sum F_{ni}(x_i\beta(\tau) - \varepsilon), \quad b_n(\varepsilon) = \frac{1}{n}\sum F_{ni}(x_i\beta(\tau) + \varepsilon)$
   et F_{ni} est la fonction de répartition de $\mathbf{Y}$ conditionnellement à x_i.

2. $\exists d > 0$ tel que $\liminf_{n\to\infty} \inf_{\|u\|=1} n^{-1}\sum \mathbb{1}(x_i.u < d) = 0.$

3. $\exists D > 0$ tel que $\limsup_{n\to\infty} \sup_{\|u\|=1} n^{-1} \sum (x_i.u)^2 \leq D$.

Remarque 2.12

La première condition permet de contrôler le comportement de la fonction de répartition au voisinage du quantile considéré. En particulier, elle est requise pour la faible consistance dans le cas univarié [97] (toujours INID). La seconde condition permet de nous assurer que les observations ne se regroupent pas préférentiellement sur un sous-espace vectoriel de $\mathbb{R}^p$.
La condition 3 contrôle la croissance de X. Elle est satisfaite, notamment lorsque l'on impose la convergence de $\frac{1}{n}X^\top X$ vers une matrice symétrique définie positive.

Théorème 2.8 (Consistance)
Les conditions 1 à 3 sont nécessaires et suffisantes à la consistance de $\widehat{\beta}(\tau)$ i.e.
$\forall \varepsilon > 0, \qquad \lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{\beta_n}(\tau) - \beta(\tau)\right| \geq \varepsilon\right) = 0.$

Vitesse de convergence vers une loi Gaussienne

Dans ce paragraphe nous supposerons, de plus, que les quantiles, conditionnellement à X, varient de façon linéaire (autrement dit les variables $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ suivent le modèle de régression quantiles usuel) $Q_{\mathbf{Y}|x_i}(\tau) = x_i\beta(\tau)$ dans un voisinage du quantile considéré.

D'autre part, on impose les conditions de régularité suivantes :

1. Les fonctions de répartitions $F_1, \ldots, F_n$ (associées aux variables $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$) sont absolument continues et admettent une densité f_i au voisinage du quantile considéré telles que $f_i(Q_{\mathbf{Y}_i}(\tau)) \neq 0$.

2. Il existe deux matrices définies positives D_0 et $D_1(\tau)$ telles que

   (a) $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n x_i^\top x_i = D_0$;

   (b) $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n f_i(Q_{\mathbf{Y}}(\tau)).x_i^\top x_i = D_1(\tau)$;

   (c) $\lim_{n\to\infty} \frac{\max_{i\in[\![1,n]\!]} \|x_i\|}{\sqrt{n}} = 0.$

Remarque 2.13

La convergence est fortement dépendante du comportement de la densité (comme le montre

[69]). Les conditions 2(a) et 2(c) sont habituellement imposées dans la littérature sur les M-estimateurs ([73] p120).

Théorème 2.9 ([103])
Pour $\tau \in]0,1[$, et sous les hypothèses de régularité de la distribution développées dans le paragraphe précédent, $\hat{\beta}_\tau$ est asymptotiquement normal :

$$\sqrt{n}(\hat{\beta}_\tau - \beta(\tau)) \xrightarrow{\mathcal{L}} N(0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1}).$$

Preuve:
Pour obtenir le comportement asymptotique de $\sqrt{n}(\hat{\beta}_\tau - \beta(\tau))$, on étudie les variations de la fonction coût autour de sa valeur optimale $(u_i)_{i\in[\![1,n]\!]} = u = \mathbf{Y} - X\beta(\tau)$:

$$\begin{aligned} Z_n &: \mathbb{R}^p \to \mathbb{R} \\ &\delta \mapsto \sum_{i=1}^{n} \left(\rho_\tau(u_i - \frac{x_i\delta}{\sqrt{n}}) - \rho_\tau(u_i)\right). \end{aligned}$$

Clairement Z_n est convexe et admet pour minimiseur $\widehat{\delta}_n = \sqrt{n}(\hat{\beta}_\tau - \beta(\tau))$
Décomposons suivant l'identité due à Knight [69] :
$\rho_\tau(u-v) - \rho_\tau(u) = -v(\tau - \mathbb{1}(u<0)) + \int_0^v \mathbb{1}(u \le s) - \mathbb{1}(u \le 0).ds$

$$\text{Alors } Z_n(\delta) = \underbrace{\sum_{i=1}^{n} -\frac{x_i\delta}{\sqrt{n}}(\tau - \mathbb{1}(u_i \le 0))}_{Z_n^{(1)}} + \underbrace{\sum_{i=1}^{n} \int_0^{\frac{x_i\delta}{\sqrt{n}}} \mathbb{1}(u_i \le s) - \mathbb{1}(u_i \le 0).ds.}_{Z_n^{(2)} = \sum_{i=1}^{n} Z_{n,i}^{(2)}}$$

- **Étude de $Z_n^{(1)}$**

  Par le théorème de Lindeberg-Feller et la condition 2(a) :
  $Z_n^{(1)} \xrightarrow{\mathcal{L}} -\delta W$ où $W \sim N(0, \tau(1-\tau)D_0)$.

- **Étude de $Z_n^{(2)}$**

$$\begin{aligned}
\mathbb{E}\left[Z_n^{(2)}\right] &= \sum_{i=1}^{n} \mathbb{E}\left[Z_{n,i}^{(2)}\right] \\
&= \sum_{i=1}^{n} \int_0^{\frac{x_i\delta}{\sqrt{n}}} F_i(u_i+s) - F_i(u_i).ds \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{x_i\delta} F_i(u_i+\frac{t}{\sqrt{n}}) - F_i(u_i).dt \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_0^{x_i\delta} \sqrt{n}(F_i(u_i+\frac{t}{\sqrt{n}}) - F_i(u_i)).dt \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_0^{x_i\delta} tf(u_i).dt + o(1) \\
&= \frac{1}{2n} \sum_{i=1}^{n} \delta^\top x_i^\top x_i \delta f(u_i) + o(1) \xrightarrow[n\to\infty]{} \frac{1}{2}\delta^\top D_1(\tau)\delta \text{ par la condition 2(b).}
\end{aligned}$$

D'autre part, on a la borne :

$$\begin{aligned}
Var(Z_n^{(2)}) &\leq \sum_{i=1}^{n} Var(Z_{n,i}^{(2)}) \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[{Z_{n,i}^{(2)}}^2\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[\int_0^{\frac{x_i\delta}{\sqrt{n}}} \mathbb{1}(u_i \leq s) - \mathbb{1}(u_i \leq 0).ds \int_0^{\frac{x_i\delta}{\sqrt{n}}} 1.ds\right] \\
&\leq \frac{1}{\sqrt{n}}|x_i\delta|\mathbb{E}\left[Z_n^{(2)}\right] \xrightarrow[n\to\infty]{} 0.
\end{aligned}$$

Puis 1(c) implique $Z_n(\delta) \xrightarrow{\mathcal{L}} Z(\delta) \overset{\text{def}}{=} -\delta W + \frac{1}{2}\delta^\top D_1(\tau)\delta$.

Z(.) est alors convexe comme limite de fonctions convexes et admet donc un unique minimum. Dans ce cadre on démontre dans [60] que :

$$\operatorname{argmin} Z_n(\delta) \xrightarrow{\mathcal{L}} \operatorname{argmin} Z(\delta) = D_1^{-1}W.$$

■

Remarque 2.14 • *Dans le cas IID, il vient* $\sqrt{n}(\hat{\beta}_\tau - \beta(\tau)) \xrightarrow{\mathcal{L}} N(0, \frac{\tau(1-\tau)}{f(Q_{\mathbf{Y}}(\tau))^2} D_0^{-1})$.

- *Cependant, il est à noter que l'estimation de la matrice de covariance n'est pas toujours satisfaisante, la variance de l'estimateur étant dépendante de la densité au quantile considéré. L'inférence peut aussi être faite à l'aide de statistiques de rang ou de méthodes bootstrap [71].*

- *Dans le cadre d'un modèle de régression non-linéaire* $Q_{\mathbf{Y}|x_i}(\tau) = g(x_i, \beta(\tau))$ *([73] p124) on obtient des résultats analogues. En effet, les hypothèses pour un tel résultat sont très similaires, la condition 1 reste inchangée alors que la condition 2 est adaptée au cadre non-linéaire en portant, non plus sur les* x_i *mais sur les* $\frac{\partial g(x_i,\beta)}{\partial \beta}|_{\beta(\tau)}$ *(ce qui prolonge le cas linéaire). Enfin, on demande un contrôle des variations moyennes de g :*
  $\exists k_1, k_2 \in \mathbb{R}^+, n_0 \in \mathbb{N}$ *tels que* $\forall \beta_1, \beta_2$,
  $k_1 \|\beta_1 - \beta_2\| \leq (\frac{1}{n}\sum_{i=1}^n (g(x_i, \beta_1) - g(x_i, \beta_2))^2)^{\frac{1}{2}} \leq k_2 \|\beta_1 - \beta_2\|$.

- *En supposant un biais modèle* $\delta(x_i)$ *(les quantiles ne s'écrivent pas comme une fonction linéaire de x) tel que* $\frac{1}{n}\sum_{i=1}^n \delta(x_i)\delta(x_i)^\top \xrightarrow[n\to\infty]{} L$ *alors* $\sqrt{n}(\hat{\beta}_\tau - \beta(\tau))$ *converge en loi vers une loi Normale.*

2.2.5 Estimation de la loi de $\hat{\beta}_\tau$

Étude de la "sparsity function"

Le principal verrou technique au théorème 2.9 est dû à l'estimation de la variance, plus particulièrement au terme correspondant à l'inverse de la densité évaluée au quantile considéré $S(\tau) = (f(Q_{\mathbf{Y}}(\tau)))^{-1}$. En effet, par définition du problème, la densité et le quantile τ sont des valeurs que nous cherchons à estimer et ne sont donc pas à disposition. Cependant, il existe différents estimateurs de la fonction S. Une des approches les plus utilisées est celle de Siddiqui (1960) que nous nous contenterons de motiver dans le cas IID :
Soit $\mathbf{Y}$ une variable aléatoire admettant une densité en $Q_{\mathbf{Y}}(\tau)$ alors
$F_{\mathbf{Y}}(Q_{\mathbf{Y}}(\tau)) = \tau \Rightarrow (Q_{\mathbf{Y}}(\tau))' = S(\tau) = (f(Q_{\mathbf{Y}}(\tau)))^{-1}$.
Il est donc naturel d'estimer S par différences finies de son estimateur $\widehat{Q_{\mathbf{Y}}}(.)$:
$\widehat{S}(\tau) = \frac{\widehat{Q_{\mathbf{Y}}}(\tau+h) - \widehat{Q}_{\mathbf{Y}}(\tau-h)}{2h}$ où $h_n \xrightarrow[n\to\infty]{} 0$ et n la taille de l'échantillon.

[17] donne, sous conditions de régularité de la FDR, une valeur optimale pour h_n au sens de l'erreur quadratique moyenne :
$h_n = n^{-\frac{1}{5}}\left(\frac{4.5S^2(\tau)}{(S''(\tau))^2}\right)^{\frac{1}{5}}$. Fort heureusement $\frac{S(\tau)}{(S''(\tau))}$ est peu sensible à la distribution de Y [75] et la pratique est de calculer h_n en imposant une densité connue.
Un autre choix de h_n dans le but d'obtenir de meilleurs intervalles de confiance est proposé par Hall et Sheater [51] : $h_n = n^{-\frac{1}{3}} z_{\alpha/2}^{\frac{2}{3}}\left(1.5\frac{S(\tau)}{(S''(\tau))}\right)^{\frac{1}{3}}$ où $z_{\alpha/2}$ est le quantile $1-\alpha/2$ de la loi normale centrée réduite.

Une fois l'expression de h_n établie, il reste à déterminer le terme $\widehat{Q_{\mathbf{Y}}}(\tau+h) - \widehat{Q_{\mathbf{Y}}}(\tau-h)$. Pour ce faire, deux choix s'offrent à nous :

- On calcule sur les résidus non-nuls $y_i - X_i\widehat{\beta}(\tau), i \in [\![1,n]\!]$ la fonction quantile empirique que l'on injecte directement dans la différence finie.
- On calcule l'intégralité du processus de régression quantile $\widehat{\beta}(.)$ et on utilise $Q_{\mathbf{Y}} = \bar{x}\widehat{\beta}(.)$. $Q_{\mathbf{Y}}$ est alors toujours croissante et a l'avantage d'offrir un estimateur consistant [104].

Les cas non-indépendants utilisent la même philosophie pour obtenir une estimation de $D_1(\tau)$ (voir théorème 2.9).

Estimation de la variance à l'aide du bootstrap

Le bootstrap est une méthode de ré-échantillonnage qui a connu un considérable essor depuis son introduction en 1979 par Efron [37]. Rappelons ici le principe général : Soit un échantillon $E = \left\{(X_1,Y_1),\ldots,(X_n,Y_n)\right\} = \left\{Z_1,\ldots,Z_n\right\}$ issue de variables supposées IID. L'idée est alors de simuler $B \in \mathbb{N}$ nouveaux échantillons de taille n en tirant avec remise parmi E, ainsi nous aurons un ensemble de B estimateurs $\{\widehat{\beta}^b(\tau) | b \in [\![1,B]\!]\}$ pour construire l'inférence. Autrement dit, nous tirons uniformément dans la fonction de quantile empirique et c'est l'idée centrale liée à ce type de méthodes.
Par exemple, considérons $\mathbf{Y}_1,\ldots,\mathbf{Y}_n$ variables aléatoires IID de fonction de répartition F et $U \sim \mathcal{U}(0,1)$ alors $F^{-1}_{\mathbf{Y}}(U) = Q_{\mathbf{Y}}(U)$ est de même loi que Y_i. Hélas, nous n'avons pas accès à $F^{-1}_{\mathbf{Y}}$ mais nous pouvons utiliser une approche "plug-in" en remplaçant $F^{-1}_{\mathbf{Y}}$ par son estimateur empirique $\widehat{F}^{-1}_{\mathbf{Y}}$ ce qui revient à tirer uniformément parmi les $Y_1,\ldots,Y_n$.
Dans la suite, nous allons énumérer les différents types de bootstrap utilisés pour estimer la

variance de $\widehat{\beta}(\tau)$. Les échantillons bootstrap seront basés sur $E = \Big\{(X_1, Y_1), \ldots, (X_n, Y_n)\Big\} = \Big\{Z_1, \ldots, Z_n\Big\}$ et notés E^b, $b \in [\![1, B]\!]$, les estimations de $\beta(\tau)$ à partir de ces derniers seront notées $\beta^b(\tau)$. La matrice de covariance Ω sera alors estimée à partir de la formule suivante :

$$\widehat{\Omega} = \frac{1}{B}\sum_{b=1}^{B}(\beta^b(\tau) - \bar{\beta}(\tau))(\beta^b(\tau) - \bar{\beta}(\tau))^\top \tag{2.9}$$

où $\bar{\beta}(\tau) = \frac{1}{B}\sum_{b=1}^{B}\beta^b(\tau)$ est la moyenne empirique prise sur les échantillons bootstrap.

- Le **"bootstrap sur les résidus"** réchantillonne les résidus de la régression quantile. Ce type de bootstrap nécessite que les erreurs soient IID.
  - On estime $\widehat{\beta}$ (en utilisant l'intégralité de l'échantillon E)
  - On tire uniformément avec remise B échantillons de taille n $\{r^b, b \in [\![1, B]\!]\}$ parmi les résidus $r = Y - X\widehat{\beta}$ pour former les échantillons bootstrap $E^b = (X_1, X_i\widehat{\beta} + r_1^b), \ldots, (X_n, X_i\widehat{\beta} + r_n^b)$
- Le **"bootstrap xy"**[75] tire avec remise uniformément parmi toutes les observations de l'échantillon : $\forall b \in [\![1, B]\!] \quad E^b = \{Z_{U_1}, , Z_{U_n}\}$ où $\forall i \in [\![1, n]\!] \quad U_i \sim \mathcal{U}_{[\![1,B]\!]}$
- L'approche de **Parzen, Wei et Ying ("pwy")**[101] est basée sur l'étude de la fonction

$$\begin{aligned} S &: \mathbb{R}^p \to \mathbb{R} \\ &\beta \mapsto \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i(\tau - \mathbb{1}(y_i - x_i\beta < 0)). \end{aligned}$$

  La distribution de $S(\beta)$ peut alors être générée par U, un vecteur aléatoire constitué de n variables de Bernoulli indépendantes de probabilité de succès τ i.e. $\forall i \in [\![1, n]\!], \quad U_i \sim \mathcal{B}(\tau)$ et $S(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i(\tau - U_i)$ l'idée est alors de générer les β^b en résolvant $S(\beta) = u^b$ où u^b est un tirage de n lois de Bernoulli. Parzen, Wei et Ying montrent que, asymptotiquement la distribution de $\widehat{\beta}(\tau) - \beta(\tau)$ peut être approximée par celle de $\tilde{\beta}(\tau) - \beta^U(\tau)$ où $\tilde{\beta}(\tau)$ est une réalisation de $\widehat{\beta}(\tau)$ et $\beta^U(\tau)$ est un vecteur aléatoire tel que $S(\beta^U) \sim U$. Cette approche permet également de gérer les cas INID.
- **Markov Chain Marginal Bootstrap ("mcmb")**

Cette technique développée tout d'abord par He et Hu (2002) [57][72] se base, elle aussi, sur un ré-échantillonnage des valeurs prises par la fonction S,
$S(\beta\widehat{(\tau)}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i z_i$ où $z_i = (\tau - \mathbb{1}(y_i - x_i\widehat{\beta}(\tau) < 0))$

1. On initialise $\beta^{(0)}(\tau) = \widehat{\beta}(\tau)$ et $k = 1$
2. on tire p échantillons bootstrap de $\{z_1, \dots, z_n\}$:
   $\forall j \in [\![1, p]\!] \quad \{z_{1,j}, \dots, z_{n,j}\}$
3. On résout les p équations :
   $\forall j \in [\![1, p]\!] \quad S_j(\beta_1^{(k)}(\tau), \dots, \beta_j^{(k)}(\tau), \beta_{j+1}^{(k-1)}(\tau), \dots) = S_j^{(k)}$
   où $S^{(k)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i z_{i,j}$ et $S_j, S_j^{(k)}$ sont les $j^{ème}$ composantes de $S, S^{(k)}$
4. On itère les étapes 2 et 3 B fois.

Les vecteurs $\beta^{(1)}, \dots, \beta^{(B)}$ forment alors, par construction, une chaîne de Markov. Par suite, asymptotiquement $\{\beta^{(i)} - \widehat{\beta}(\tau), i \in [\![1, B]\!]\}$ est un processus $AR(1)$ (multivarié) à partir duquel, nous pouvons dériver un estimateur consistant de la matrice de variance-covariance de $\widehat{\beta}(\tau)$. Cette méthode a plusieurs vertus. Elle permet de faciliter l'approche bootstrap en grande dimension paramétrique, en effet elle ne nécessite pas la résolution d'un système de dimension p mais la résolution de p équations de dimension 1. De plus, elle peut être adaptée au cas INID à l'aide de transformations simples (MCMB-AB) [70].

- **Block bootstrap** est une approche très utilisée dans l'étude des séries temporelles. Elle permet notamment de pallier la corrélation temporelle entre les Y_i (en effet les méthodes précédentes ne prennent pas en compte cette structure dans l'inférence).

  – **Simple Block bootstrap**

  1. On partitionne $E = \sqcup_{k=1}^{K} Bl_k$ en K blocs d'observations consécutives de taille $Bl_k \approx \frac{n}{K}$.
  2. On tire uniformément avec remise parmi les blocs pour former de nouveaux échantillons bootstrap :
     $\forall b \in [\![1, B]\!] \quad E^b = \{Bl_{U_1}, , Bl_{U_K}\}$ où $\forall i \in [\![1, B]\!] \quad U_i \sim \mathcal{U}_{[\![1,B]\!]}$.

  – **Moving Blocs bootstrap[88]**

1. On définit les blocs d'observations consécutives de taille l :
   $\forall i \in [\![1, n-l+1]\!] \quad Bl_i = \{Z_i, , Z_{i+l-1}\}$.
2. On tire uniformément avec remise $k = [\frac{n}{l}]$ blocs pour former les nouveaux échantillons bootstrap :
   $\forall b \in [\![1, B]\!] \quad E^b = \{Bl_{U_1}, , Bl_{U_K}\}$ où $\forall i \in [\![1, B]\!] \quad U_i \sim \mathcal{U}_{[\![1,n-l+1]\!]}$.

Pour l'estimation de la matrice de variance-covariance de $\hat{\beta}(\tau)$, il existe des tailles de bloc asymptotiquement optimales (au sens de l'erreur quadratique) dans le cas autocorrélé : $l_{op} = (1.5n)^{\frac{1}{3}} \zeta^{\frac{-2}{3}}$ où $\zeta = \frac{\gamma(0)+2\sum_{i=1}^{\infty}\gamma(i)}{|\sum_{i=1}^{\infty} i\gamma(i)|}$ où $\gamma(i) = cov(Y_1, Y_{i+1})$

2.2.6 Suppléments méthodologiques

Séries temporelles

La régression quantile s'adapte aussi aux modèles usuels associés à l'étude de séries temporelles. Elle permet notamment de généraliser les modèles auto-régressifs. En effet, [81] propose de se placer dans le modèle suivant (QAR) :

$$Q_{Y_t}(\tau) = \beta_0(\tau) + \sum_{i=1}^{k} \beta_i(\tau) Y_{t-i}.$$

On revient donc au modèle auto-régressif habituel en imposant aux β_i d'être constants. Bien sûr, d'autres généralisations sont possibles [76] p11. Ce type de modèle n'est pas implémenté durant cette thèse mais permettrait d'explorer la dépendance temporelle des séries d'observations.

Méthodes de régularisation

Les méthodes de régularisation développées dans la section précédente à l'aide de la théorie des RKHS peuvent, bien sûr, être mises à profit pour la régression quantile. Cependant, le problème d'optimisation, n'étant plus forcément un problème linéaire, il peut devenir plus ardu, c'est-à-dire plus long à résoudre concernant le temps de calcul. C'est le cas des splines cubiques de lissages sur l'intervalle $[0, 1]$. Pour contourner cette difficulté, on peut considérer

d'autres types de régularisations [79] comme les normes L^p appliquées à la dérivée seconde :

$$\underset{s\in\mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau\big(Y_i - s(x_i)\big) + \lambda \|s''\|_{L^p}$$

où

- $p \geq 1$
- $0 < x_1 < \cdots < x_n < 1$
- $\mathcal{F}$ un espace fonctionnel .

Le cas p=2 (problème d'optimisation quadratique) pose de sérieux problèmes de temps de calcul, du moins avec les algorithmes proposés [20]. D'un autre côté, les cas $p = 1$ et $p = \infty$ sont particulièrement attractifs car ils peuvent être ré-écrits sous la forme d'un programme linéaire. L'étude du cas $p = 1$ mènera à considérer des fonctions linéaires par morceaux. Celle de $p = \infty$ admettra pour minimiseurs, des splines quadratiques [79].

Théorie du ré-arrangement

Lorsque plusieurs quantiles sont estimés séparément, il peut arriver que ceux-ci se croisent. En effet, aucune condition ne nous assure le contraire comme le montre l'exemple (fait à partir de données synthétiques) de la figure 2.2. Ce problème a donné lieu à la théorie des ré-arrangements, qui permet de ré-ordonner les quantiles estimés à posteriori [27].

Pour se faire, nous raisonnons à une valeur fixée des covariables x et depuis une estimation de la fonction quantile $\hat{Q}(.|x)$. Pour obtenir une telle fonction à partir de N_α quantiles de niveaux (croissants) $\{\tau_i\}_{i\in[\![1,N_\alpha]\!]}$, nous pouvons, par exemple, utiliser sa fonction de quantile empirique :
$\hat{Q}(u|x) \overset{\text{def}}{=} \hat{Q}(\tau_1|x)\mathbb{1}(u \in [0,\tau_1]) + \sum_{i=2}^{N_\alpha} \hat{Q}(\tau_i|x)\mathbb{1}(u \in [\tau_{i-1},\tau_i])$ où $u \in]0,1[$. Si $\hat{Q}(.|x)$ est un estimateur non-croissant en u, pour chaque x il existe une façon de réarranger $\hat{Q}(.|x)$ qui produit une fonction monotone $u \to \hat{Q}^*(u|x)$. La méthode remplace les quantiles initiaux par les quantiles de la variable aléatoire $\mathbf{Y}_x := \hat{Q}(\mathbf{U}|x)$ où $\mathbf{U} \sim \mathcal{U}(0,1)$. Plus précisément :

$\hat{F}(y|x) \overset{\text{def}}{=} \mathbb{P}(\mathbf{Y}_x < y) = \mathbb{P}(\hat{Q}(\mathbf{U}|x) < y) = \int_0^1 \mathbb{1}(\hat{Q}(\mathbf{U}|x) < y).du$
$\hat{Q}^*(u|x) \overset{\text{def}}{=} \hat{F}^{-1}(u|x) = \inf\{y|\hat{F}(y|x) \geq u\}.$

Remarque 2.15

Lorsque nous considérons un ensemble régulièrement espacé de quantiles $\tau \in \{\frac{1}{N_\alpha+1}, \ldots, \frac{N_\alpha}{N_\alpha+1}\}$, *alors en notant les quantiles estimés* $Q_i \overset{\text{def}}{=} \hat{Q}(\frac{i}{N_\alpha+1}|x)$ *pour* $i \in [\![1, N_\alpha]\!]$, *le ré-arrangement est équivalent à ré-ordonner les quantiles pour* x *fixé :*
En effet, $\hat{F}(y|x) = \dfrac{card(\{Q_i|Q_i < y\}}{N_\alpha}$. *Par suite,* $\hat{Q}^*(\frac{i}{N_\alpha}|x) = Q_{(i)}$ *où* $Q_{(i)}$ *est la* $i^{\text{ème}}$ *plus petite valeur (la statistique d'ordre de rang* i*).*

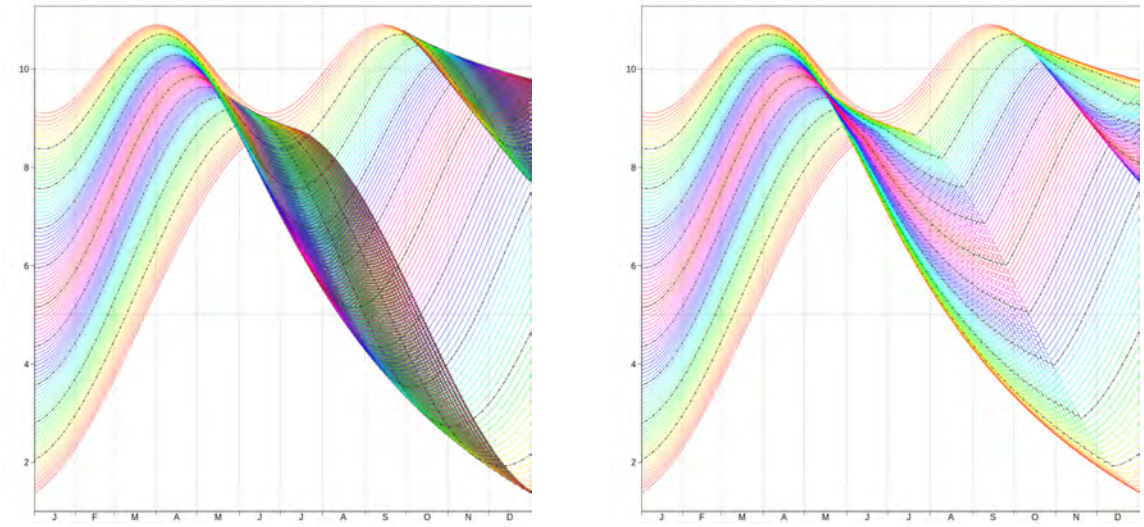


FIGURE 2.2 – Dans cet exemple fait à partir de données synthétiques, chaque courbe représente l'estimation d'un centile de la variable d'intérêt le long de l'année. Les quantiles sont décomposés dans la base de Fourier discrète de période 365. On peut constater de sérieux problèmes de croisement, d'une part à cause du passage de cycles annuels uni-modaux (pour les bas quantiles) à bi-modaux (pour les hauts quantiles) et d'autre part, dus à une réduction de la variance de la variable d'intérêt en juin (sur la figure de gauche). Le ré-arrangement des courbes quantiles peut être vu sur la figure de droite.

Bien sûr, il existe des développements de ces méthodes pour réarranger les quantiles de façon lisse [34], mais par soucis de concision, nous préférons décrire le concept général qui est suffisant pour les besoins de cette thèse.

2.3 Sélection de modèles et erreur de généralisation

2.3.1 Cadre et premières définitions

On cherche à expliquer une variable aléatoire Y, du moins une de ses caractéristiques (e.g telle que son espérance conditionnelle), par un vecteur aléatoire X de $\mathbb{R}^p$. Pour cela, nous avons à notre disposition un échantillon $T \overset{\text{def}}{=} \left\{(X_1, Y_1), \ldots, (X_n, Y_n)\right\}$ issu de la même loi mère F, $\hat{f}$ un estimateur entraîné sur T et L une fonction coût permettant de quantifier l'erreur commise entre $\hat{f}(X)$ et Y : $L(\hat{f}(X), Y)$ où (Y, X) sont une nouvelle réalisation de F.

Exemple 2.2
Pour ce qui est de la régression, nous avons déjà vu deux fonctions coûts : l'une associée à la moyenne (coût quadratique)

$$\begin{aligned} L &: \mathbb{R}^2 \to \mathbb{R} \\ (a, b) &\mapsto (a-b)^2 \end{aligned}$$

et l'autre aux quantiles

$$\begin{aligned} L &: \mathbb{R}^2 \to \mathbb{R} \\ (a, b) &\mapsto (a-b)(\tau - \mathbb{1}(a-b<0)). \end{aligned}$$

Pour ce qui est de la classification supervisée (Y à valeurs dans $\mathcal{G} = [\![1, K]\!]$, $K \in \mathbb{N}^*$) les fonctions coûts de loin les plus utilisées sont les pertes $0-1$ et la vraisemblance

$$\begin{aligned} L &: \mathbb{R}^2 \to \mathbb{R} \\ (a, b) &\mapsto \mathbb{1}(a \neq b) \end{aligned} \qquad \begin{aligned} L &: \mathbb{R}^2 \to \mathbb{R} \\ (Y, \widehat{p}(X)) &\mapsto -\log(\Pi_{k=1}^{K} \widehat{p}_k^{(\mathbb{1}_{(Y=k)})}(X)) \end{aligned}$$

où $\widehat{p}_k$ représente un estimateur de la probabilité d'appartenir au groupe k (conditionnellement à X).

Pour quantifier les erreurs, nous nous intéresserons principalement à trois quantités.

Définition 2.4 (Erreur d'entraînement, de test et de généralisation)

- *Erreur de généralisation :* C'est une variable aléatoire représentant l'erreur de prévision conditionnellement à son échantillon d'apprentissage T : $Err_T \overset{\text{def}}{=} \mathbb{E}\left[L(\hat{f}(X), Y)|T\right]$.
- *Erreur de test :* Il s'agit de l'espérance de l'erreur de généralisation : $Err \overset{\text{def}}{=} \mathbb{E}\left[Err_T\right] = \mathbb{E}\left[L(\hat{f}(X), Y)\right]$.
- *Erreur d'entraînement :* $\bar{err} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} L(\hat{f}(X_i), Y_i)$.

L'étude de l'erreur de généralisation nous apportera, d'une part, des informations sur les capacités prédictives des modèles étudiés mais aussi des méthodes de sélection de modèle. Par exemple, lorsque l'estimation de la complexité du modèle est nécessaire. Ces deux composantes seront exploitées autant dans le cas des normales que dans celui des quantiles. Malheureusement, l'on est rarement dans une situation où l'on peut obtenir de "bons" estimateurs de Err_T, nous nous contenterons donc d'étudier son espérance Err. Hélas, l'erreur d'entraînement n'est pas un bon estimateur de l'erreur de test comme le montre la figure 2.3, il fournit une estimation trop optimiste en général.

2.3.2 Décomposition biais-variance pour un coût quadratique

Nous nous plaçons dans le cadre du modèle suivant :

$$Y = f(X) + \varepsilon \tag{2.10}$$

où $\mathbb{E}\left[\varepsilon\right] = 0$ et $var(\varepsilon_i) = \sigma^2$.

Théorème 2.10 (Décomposition biais-variance)
Soit $x_0 \in \mathbb{R}^p$ alors, on a la décomposition suivante :

$$\begin{aligned} Err(x_0) &= \mathbb{E}\left[(Y - \widehat{f}(x_0))^2 | X = x_0\right] && (2.11)\\ &= \sigma^2 + \left(\mathbb{E}\left[\widehat{f}(x_0)\right] - f(x_0)\right)^2 + \mathbb{E}\left[\left(\widehat{f}(x_0) - \mathbb{E}\left[\widehat{f}(x_0)\right]\right)^2\right] && (2.12)\\ &= \textit{Erreur Irréductible} + \textit{biais}(\widehat{f}(x_0))^2 + var(\widehat{f}(x_0)). && (2.13) \end{aligned}$$

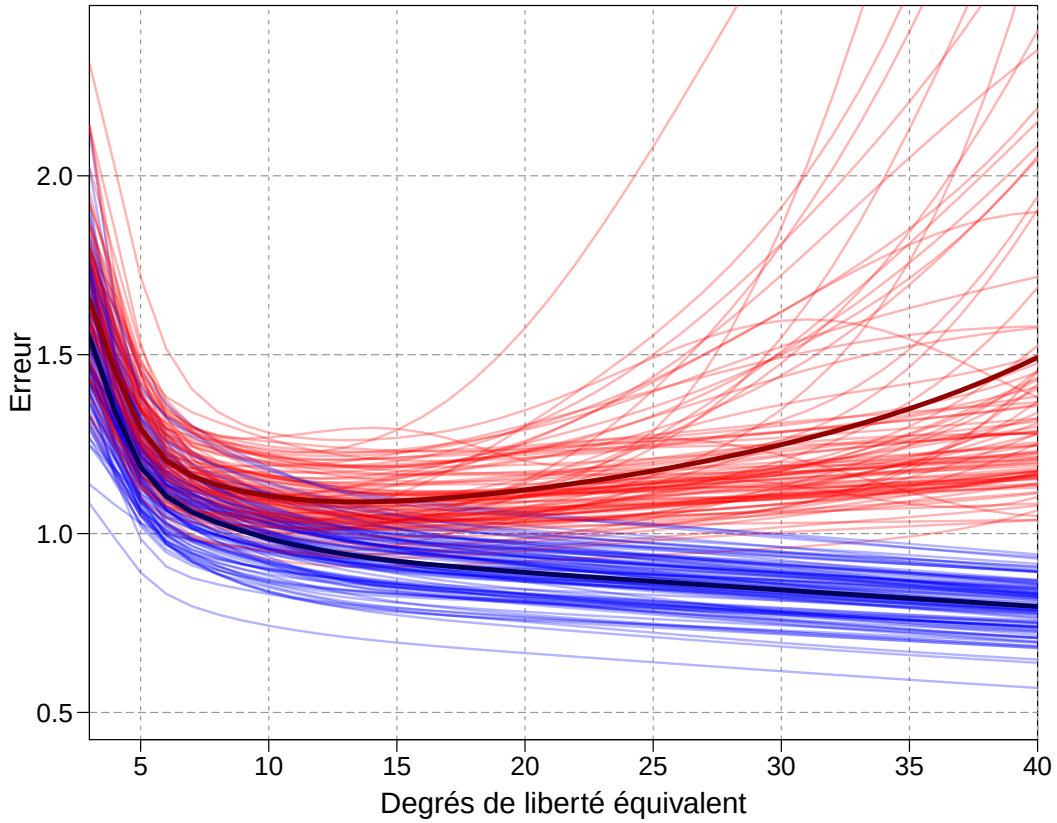


FIGURE 2.3 – Comportement des erreurs en fonction de la complexité de l'estimateur : Les courbes rouges claires montrent plusieurs réalisations de l'erreur de généralisation Err_T pour 100 échantillons d'apprentissage de taille 200. La courbe rouge foncé montre l'erreur de test Err. De même, les courbes bleues montrent plusieurs réalisations de l'erreur d'entraînement $\bar{err}$ sur les mêmes échantillons et leurs espérances sur T. Les échantillons sont issus du modèle suivant $Y = f(X) + \varepsilon$ où X, ε suivent des lois normales centrées réduites et f est une spline. Nous avons pris pour estimateur $\hat{f}_\lambda$ une spline de lissage. La complexité du modèle est alors représentée en terme de degrés de liberté équivalents.

Preuve:

$$
\begin{aligned}
Err(x_0) &= \mathbb{E}\left[(Y - \widehat{f}(x_0))^2 | X = x_0\right] \\
&= \mathbb{E}\left[((Y - f(x_0)) + (f(x_0) - \mathbb{E}\left[\widehat{f}(x_0)\right]) + (\mathbb{E}\left[\widehat{f}(x_0)\right] - \widehat{f}(x_0)))^2 | X = x_0\right] \\
&= \mathbb{E}_T\left[\mathbb{E}_Y\left[((Y - f(x_0)) + (f(x_0) - \mathbb{E}\left[\widehat{f}(x_0)\right]) + (\mathbb{E}\left[\widehat{f}(x_0)\right] - \widehat{f}(x_0)))^2 | X = x_0\right]\right]
\end{aligned}
$$

car Y et T sont indépendants.

Il suffit de montrer que l'espérance des termes croisés est nulle :

- $\mathbb{E}_Y\left[(Y - f(x_0)).(f(x_0) - \mathbb{E}\left[\widehat{f}(x_0)\right])|X = x_0\right]$
  $= \mathbb{E}_Y\left[(Y - f(x_0))|X = x_0\right].(f(x_0) - \mathbb{E}\left[\widehat{f}(x_0)\right]) = 0$ car $\mathbb{E}_Y[Y] = f(x_0)$.
- De même,
  $\mathbb{E}_Y\left[(Y - f(x_0)).(\mathbb{E}\left[\widehat{f}(x_0)\right] - \widehat{f}(x_0))|X = x_0\right] = \mathbb{E}_Y\left[(Y - f(x_0))|X = x_0\right].(\mathbb{E}\left[\widehat{f}(x_0)\right] - \widehat{f}(x_0)) = 0$
- $\mathbb{E}_T\left[(f(x_0) - \mathbb{E}_T\left[\widehat{f}(x_0)\right]).(\mathbb{E}_T\left[\widehat{f}(x_0)\right] - \widehat{f}(x_0))\right] = \mathbb{E}_T\left[(\mathbb{E}_T\left[\widehat{f}(x_0)\right] - \widehat{f}(x_0))\right].(f(x_0) - \mathbb{E}_T\left[\widehat{f}(x_0)\right]) = 0.$

■

Remarque 2.16

- Dans ce modèle $Err(x_0)$ est minorée par σ^2 l'erreur irréductible inhérente au modèle.
- En général, lorsque la complexité de $\widehat{f}$ augmente on ne peut gagner sur les deux tableaux le biais et la variance de l'estimateur. En effet, dans le cadre de la sélection d'un modèle prédictif, le terme de biais correspond à l'erreur commise en supprimant ou en pénalisant trop certaines variables du vrai modèle, le terme de variance mesure l'incertitude liée à l'estimation des paramètres du modèle.

Exemple 2.3

Voici quelques exemples permettant de corroborer cette dernière remarque sur le compromis biais/variance :

- *K-plus proches voisins :* $\widehat{f}_k$.
  Ici la complexité de l'estimateur est contrôlée par le nombre voisins k. $Err(x_0) = \sigma^2 + (f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x^0_{(l)}))^2 + \frac{1}{k}\sigma^2$ où $x_{(l)}$ représente les plus proches voisins de x_0. En moyenne sur l'échantillon : $\frac{1}{n}\sum_{i=1}^{n} Err(x_i) = \sigma^2 + \frac{1}{n}(f(x_i) - \frac{1}{k}\sum_{l=1}^{k} f(x^i_{(l)}))^2 + \frac{1}{k}\sigma^2$
  On voit donc que lorsque k diminue (la complexité de l'estimateur augmente), la variance de $\widehat{f}_k$ augmente et si la fonction f est suffisamment régulière, le biais diminue et inversement.

- *Régression linéaire :* $\widehat{f}_p$.
  $\widehat{f}_p(x_0) = x_0(X^\top X)^{-1}X^\top Y$
  $Err(x_0) = \sigma^2 + (f(x_0) - x_0(X^\top X)^{-1}X^\top f(X))^2 + \frac{p}{n}\sigma^2 \left\|x_0(X^\top X)^{-1}X^\top\right\|_2^2$
  En moyenne sur l'échantillon :
  $\frac{1}{n}\sum_{i=1}^n Err(x_i) = \sigma^2 + \frac{1}{n}\left\|(I_n - X(X^\top X)^{-1}X^\top)f(X)\right\|_2^2 + \frac{p}{n}\sigma^2$ De même ici lorsque p augmente la variance croit de façon linéaire.

La décomposition biais-variance ci-dessus ne se généralise pas (hélas) à toute fonctions coût et bien que d'autres formes de décomposition de l'erreur existent dans des cadres plus généraux [63, 35] à notre connaissance, il n'y a pas de consensus sur une formule générique. Nous laisserons donc là la discussion sur ce sujet.

La décomposition biais-variance, appliquée à l'erreur quadratique moyenne (MSE), sera utilisée pour étudier les estimateurs des normales dans le chapitre 3.

2.3.3 Degrés de liberté

2.3.3.1 Pour les estimateurs linéaires

Pour les estimateurs s'écrivant comme une combinaison linéaire du vecteur Y (i.e tels que $\widehat{Y} = S.Y$ où $S \in M_n(\mathbb{R})$) il existe des formules simples permettant d'obtenir la dimension paramétrique du modèle.

Proposition 2.2
Soit $\widehat{Y} = SY$ *où* $S \in M_n(\mathbb{R})$, *en supposant de plus que* $cov(Y) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ *alors* $df(\widehat{Y}) \overset{\text{def}}{=} \sum_{i=1}^n \frac{cov(\widehat{Y}_i, Y_i)}{\sigma_i^2} = tr(S)$.

Preuve:
$cov(\widehat{Y}_i, Y_i) = \mathbb{E}\left[(e_i^\top SYY^\top e_i)\right] = e_i^\top S\mathbb{E}\left[(YY^\top)\right]e_i = e_i^\top Scov(Y)e_i$
donc $\sum_{i=1}^n \frac{cov(\widehat{Y}_i, Y_i)}{\sigma_i^2} = \sum_{i=1}^n \frac{e_i^\top Scov(Y)e_i}{\sigma_i^2} = tr(S)$. ■

2.3.3.2 Pour la régression quantile

Nous avons déjà vu, dans la deuxième section de ce chapitre, que la régression quantile avec p co-variables pouvait s'écrire comme un problème linéaire dont la solution impliquait l'interpolation de p points. Nous étendrons donc, dans ce cadre, le nombre de paramètres au nombre de résidus $Y - X\widehat{\beta}_\tau$ nuls : $df \stackrel{\text{def}}{=} Card(\{y_i = X_i\widehat{\beta}_\tau | i \in [\![1, n]\!]\})$.

2.3.4 De l'optimisme à la véritable erreur

Le but de cette section est de comprendre en quoi l'erreur d'apprentissage diffère de l'erreur de généralisation en se restreignant à des valeurs fixées de la variable X. L'échantillon s'écrira donc $T = \Big\{(x_1, Y_1), \ldots, (x_n, Y_n)\Big\}$.

Définition 2.5

Lorsque nous considérons des réalisations des régresseurs $x_1, \ldots, x_n$ nous pouvons définir l'erreur d'apprentissage et l'erreur de test de manière similaire :
$\bar{err} = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \widehat{f}(x_i))$, $Err_{in} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Y_i^0}\left[L(Y_i^0, \widehat{f}(x_i))|T\right]$
où $T^0 = (x_1, Y_1^0), \ldots, (x_n, Y_n^0)$ est indépendant et de même loi que T.
On définit alors l'optimisme moyen comme la différence moyenne entre l'erreur de test et d'apprentissage : $w \stackrel{\text{def}}{=} \mathbb{E}_T[Err_{in} - \bar{err}]$.

Théorème 2.11

$\mathbb{E}_T[Err_{in}] = \mathbb{E}_T[\bar{err}] + \frac{2}{n}\sum_{i=1}^{n} cov(\widehat{Y}_i, Y_i)$ *et* $w = \frac{2}{n}\sum_{i=1}^{n} cov(\widehat{Y}_i, Y_i)$.

Preuve:

Dans le cas d'un coût quadratique

$$\begin{aligned}
w &= \frac{1}{n}\mathbb{E}_T\left[\sum_{i=1}^{n}(\mathbb{E}_{Y_i^0}\left[L(Y_i^0,\widehat{f}(x_i))|T\right] - L(Y_i,\widehat{f}(x_i)))\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{Y_i^0}\left[(Y_i^0)^2|T\right] - 2\mathbb{E}_T\left[\mathbb{E}_{Y_i^0}\left[Y_i^0.\widehat{f}(x_i)|T\right]\right] + \mathbb{E}_T\left[\widehat{f}(x_i)^2\right]\right. \\
&\qquad \left. - \mathbb{E}_T\left[(Y_i)^2\right] + 2\mathbb{E}_T\left[Y_i.\widehat{f}(x_i)\right] - \mathbb{E}_T\left[\widehat{f}(x_i)^2\right]\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}2(\mathbb{E}_T\left[Y_i.\widehat{f}(x_i)\right] - \mathbb{E}_T\left[\mathbb{E}_{Y_i^0}\left[Y_i^0.\widehat{f}(x_i)|T\right]\right]) \\
&= \frac{1}{n}\sum_{i=1}^{n}2(\mathbb{E}_T\left[Y_i.\widehat{f}(x_i)\right] - \mathbb{E}_T\left[\widehat{f}(x_i)\right]\mathbb{E}_{Y_i^0}\left[Y_i^0\right]) \\
&= \frac{2}{n}\sum_{i=1}^{n}cov(\widehat{Y}_i, Y_i)
\end{aligned}$$

$w = \frac{1}{n}\mathbb{E}_T\left[\sum_{i=1}^{n}(\mathbb{E}_{Y_i^0}\left[L(Y_i^0,\widehat{f}(x_i))|T\right] - L(Y_i,\widehat{f}(x_i)))\right]$. ■

Exemple 2.4

Dans le cas de la régression linéaire : $Y = f(X) + \varepsilon$, $\widehat{Y} = X\widehat{\beta}$, $cov(\varepsilon) = \sigma^2 I_n$

$\mathbb{E}_T[Err_{in}] = \mathbb{E}_T[e\bar{r}r] + \frac{2}{n}\sum_{i=1}^{n}cov(\widehat{y}_i, y_i) = \mathbb{E}_T[e\bar{r}r] + 2\frac{p}{n}$.

2.3.5 Critères C_p AIC BIC

Une première stratégie pour estimer l'erreur de prédiction consiste à estimer l'optimisme : $\widehat{Err}_{in} = e\bar{r}r + \widehat{w}$. Cela nous permettra de mettre en place des critères de sélection de modèles, nous ne motiverons ici que les trois plus connus :

- Cadre de la régression par moindres carrés : C_p de Mallows (1973) [94]
  $C_p \overset{\text{def}}{=} e\bar{r}r + 2\frac{df(\widehat{Y})}{n}.\widehat{\sigma^2}$ où $\widehat{\sigma^2}$ estimateur de σ^2.

- Pour les estimateurs par maximum de vraisemblance : Akaïke Information Criterion (AIC) 1974 [4]
  Elle se base sur une approximation $\mathbb{E}_Y\left[log(P_{\widehat{\theta}})\right] \approx \frac{1}{n}(\mathbb{E}_T[loglik] + df(\widehat{Y}))$
  où $loglik = \max_{\theta\in\Theta}(\sum_{i=1}^{n}(log(P_\theta(Y_i))))$, $\widehat{\theta} = \text{argmax}_{\theta\in\Theta}(\sum_{i=1}^{n}(log(P_\theta(Y_i))))$
  $AIC \overset{\text{def}}{=} \frac{-2}{n}(\mathbb{E}_T[loglik] - df(\widehat{Y}))$

Remarque 2.17

Dans le cadre du modèle linéaire gaussien avec variance de l'erreur connue, il est équivalent de minimiser le critère AIC et C_p.

- Le critère d'information bayésien (BIC/SIC) :

  Toujours pour les estimateurs du maximum de vraisemblance, le point de vue Bayesien offre une autre approche.

  Cadre : Soit $\{M_m\}_{m\in K}$ une collection de modèles et θ_m les paramètres associés au modèle M_m. On munit les lois à priori $Pr(M_m)$ et $Pr(M_m|\theta_m)$ d'une loi uniforme. On veut alors sélectionner le modèle le plus vraisemblable au vu de l'échantillon T :

$$\begin{aligned} Pr(M_m|T) &\propto Pr(M_m).Pr(T|M_m) \\ &\propto Pr(M_m).\int Pr(T|\theta_m, M_m)Pr(\theta_m|M_m)d\theta_m. \end{aligned}$$

  Si $Pr(M_m) \sim \mathcal{U}$ il suffit de maximiser $Pr(T|M_m)$ que l'on choisit d'approximer par la méthode de Laplace :
  $\log(Pr(T|M_m)) = \log(Pr(T|\widehat{\theta}_m, M_m)) - \frac{df(M_m)}{2}\log(n) + O(1)$. Par suite, on est amené à minimiser le critère suivant $BIC \stackrel{\text{def}}{=} -2loglik + log(n)df(M_m)$.

Remarque 2.18

Dans le cas du modèle Gaussien homoscédastique à variance connue σ^2 : $BIC = \frac{n}{\sigma^2}[e\bar{r}r + log(n).\frac{df}{n}\sigma^2]$ Ce critère pénalise donc plus fortement les modèles complexes (si $n > 7$) que le critère AIC.

La grande qualité de ses critères réside dans le fait qu'ils ne nécessitent qu'une seule évaluation de la vraisemblance, ce qui n'est pas le cas des méthodes de validation croisée. Offrons quelques éléments de comparaisons quant aux capacités des critères AIC et BIC. Soient $\mathcal{C} \stackrel{\text{def}}{=} M_1, \ldots, M_K$ une collection de modèles et M_* le modèle dont les données sont issues (vrai modèle),

- Le critère BIC est asymptotiquement consistant [53] i.e. il choisira, asymptotiquement, avec probabilité 1 le "vrai" modèle (celui dont les données sont issues) parmi une famille

de modèles, ce qui n'est pas le cas du critère AIC.
Autrement dit, $\text{argmin}_{M\in\mathcal{C}}(BIC(M)) \xrightarrow[n\to\infty]{\mathcal{P}} M_*$ et ce n'est pas le cas pour le critère AIC même si $M_* \in \{M_1, \ldots, M_K\}$ [18, 25, 100]

- Dans le cadre du coût quadratique et en supposant $M_* \in \{M_1, \ldots, M_K\}$ le critère AIC est optimal asymptotiquement (i.e. il choisit le modèle ayant l'erreur la plus faible), le BIC ne l'est pas [140].

Parmi les inconvénients majeurs, il y a le fait que ces critères nécessitent beaucoup de données ($> p^2$) et sont mal adaptés à la grande dimension.

Ces critères seront utilisés et comparés à la validation croisée quasi systématiquement dans le contexte de l'estimation des normales comme dans celui des distributions.

2.3.6 Validation croisée

Ici, nous nous intéressons de nouveau à obtenir des garanties sur Err. Pour ce faire, nous allons étudier une technique de validation croisée "K-fold".

2.3.6.1 K-fold cross-validation

Idée : on partitionne notre échantillon T en K parties égales, puis on entraîne l'estimateur sur tout l'échantillon privé d'une de ses parties, qui servira d'échantillon test. Ensuite on répète l'opération K fois et on calcule la moyenne des erreurs sur les K échantillons tests. Ainsi, nous obtenons une estimation de l'erreur test représentative du comportement de l'estimateur en dehors de son échantillon d'apprentissage.

Définition 2.6
Soient $k\colon [\![1,n]\!] \to [\![1,K]\!]$, $K < n$ *définissant une partition de* T; $\widehat{f}^{-k(i)}$ *estimateur entraîné sur* $[\![1,n]\!]\backslash\{k^{-1}(k(i))\}$.
Définissons la statistique K-fold par : $CV(\widehat{f}) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \widehat{f}^{-k(i)}(x_i))$. *Il est fréquent que nous considérions une suite d'estimateurs* $\widehat{f}_\alpha(x)$ *indexés par* α, *nous noterons alors :* $CV(\widehat{f}_\alpha) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \widehat{f}_\alpha^{-k(i)}(x_i))$.

Propriété 2.5 (CV vu comme estimateur de Err [43, 13])
La validation croisée K-fold founit un estimateur asymptotiquement sans biais de Err : $Biais(CV) = \frac{c_0}{n(K-1)} + o(\frac{1}{n(K-1)})$ *où* c_0 *est une constante dépendante de* L *et de la loi du couple* (X, Y).

Remarque 2.19
Lorsque le nombre de fold K *augmente, le biais diminue. Cependant il n'est pas toujours vrai que la variance augmente. Par exemple, pour le modèle linéaire Gaussien avec erreur connue de variance* σ^2 *[23] :* $var(CV) = \frac{2\sigma^2}{n} + \frac{4\sigma^4}{n^2}[4 + \frac{4}{K-1} + \frac{2}{(K-1)^2} + \frac{1}{(K-1)^3}] + o(n^{-2})$.

Il est à noter que pour obtenir $CV(\widehat{f})$, il est nécessaire d'entraîner K estimateurs de f en prenant en compte un nombre d'observations généralement comparable à celui de l'échantillon tout entier, notamment lorsque K est proche de n. Fort heureusement, pour certains modèles tels que les splines de lissage et de nombreux estimateurs linéaires (ainsi que pour le cas extrême $K = n$ "Leave-one-out"), il n'est pas indispensable de réitérer K fois l'apprentissage de l'estimateur sur autant de sous-échantillons, l'information étant contenue dans l'estimateur ajusté sur l'intégralité des données.

Théorème 2.12
Soit S_λ *matrice de lissage de l'estimateur spline (i.e.* $S_\lambda Y = \widehat{f}_\lambda$*)*
$CV(f, \lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}_\lambda^{-i}(x_i))^2 = \frac{1}{n}\sum(\frac{y_i - \widehat{f}_\lambda(x_i)}{1 - S_\lambda(i,i)})^2$.

Preuve:
Posons :

- $j \in [\![1, n]\!]$.
- $\widehat{f}^{(-j)}$ spline de lissage entraînée sur $T \backslash (x_j, y_j)$.
- $\widehat{f}^*$ spline de lissage entraînée sur $T^* \overset{\text{def}}{=} \{(x_1, y_1^*), \ldots, (x_n, y_n^*)\}$
  où $y_i^* = y_i$ si $i \neq j$ et $y_j^* = \widehat{f}^{(-j)}(x_j)$.
  Nous noterons $S_\lambda = (S_\lambda(i,j))_{(i,j)\in[\![1,n]\!]}$ sa matrice de lissage.

Autrement dit, nous nous plaçons dans le cas particulier où la valeur du nouvel échantillon

T^* en x_j est exactement la valeur prise par l'estimateur spline entraîné sur $T\backslash(x_j, y_j)$.

Comparons les valeurs prises par le problème (2.1) :

$$\begin{aligned}\sum_{i\in[\![1,n]\!]\backslash\{j\}}(y_i^*-\widehat{f}_\lambda^*(x_i))^2+\lambda\left\|f^{*''}\right\|_{L^2} &\geq \sum_{i\in[\![1,n]\!]\backslash\{j\}}(y_i^*-\widehat{f}_\lambda^{(-j)}(x_i))^2+\lambda\left\|f^{(-j)''}\right\|_{L^2}\\ &\geq \sum_{i\in[\![1,n]\!]\backslash\{j\}}(y_i-\widehat{f}_\lambda^{-j}(x_i))^2+\lambda\left\|f^{(-j)''}\right\|_{L^2}\end{aligned}$$

donc $\widehat{f}_\lambda^*(x_j)=\widehat{f}_\lambda^{(-j)}(x_j)=y_j^*$,
sinon $\sum_{i\in[\![1,n]\!]}(y_i^*-\widehat{f}_\lambda^*(x_i))^2+\lambda\left\|f^{*''}\right\|_{L^2}>\sum_{i\in[\![1,n]\!]}(y_i^*-\widehat{f}_\lambda^{(-j)}(x_i))^2+\lambda\left\|f^{(-j)''}\right\|_{L^2}$ ce qui est absurde car $\widehat{f}_\lambda^*$ est le minimiseur du problème (2.1) sur T^*.

$$\begin{aligned}y_j^*=\sum_{i=1}^n S_\lambda(j,i)y_i^* &= \sum_{i\in[\![1,n]\!]\backslash\{j\}}S_\lambda(j,i)y_i+S_\lambda(j,j)y_j^*\\ &=\sum_{i=1}^n S_\lambda(j,i)y_i^*=\sum_{i=1}^n S_\lambda(j,i)y_i+S_\lambda(j,j)(y_j^*-y_j).\end{aligned}$$

En soustrayant y_j de part et d'autre de cette équation nous obtenons

$$\begin{aligned}(1-S_\lambda(j,j))(y_j^*-y_j)&=\sum_{i=1}^n S_\lambda(j,i)y_i-y_j\\ \widehat{f}_\lambda^{(-j)}(x_j)-y_j&=\frac{\sum_{i=1}^n S_\lambda(j,i)y_i-y_j}{1-S_\lambda(j,j)}.\end{aligned}$$

■

Remarque 2.20

Soient $I = \{i_1, , i_l\} \in [\![1, n]\!]$ *et* $\widehat{f}^{-I}$ *spline de lissage entraînée sur* $T \backslash \{(x_i, y_i) | i \in I\}$ *et* $\widehat{f}$ *la spline de lissage entraînée sur l'intégralité de l'échantillon alors :*

$(\widehat{f}^{-I}(x_i) - y_i)_{i \in I} = (I_l - S_\lambda(I))^{-1} Y_I$

$$où\ I_l - S_\lambda(I) = \begin{pmatrix} 1 - S_\lambda(i_1, i_1) & -S_\lambda(i_1, i_2) & \dots & -S_\lambda(i_1, i_l) \\ -S_\lambda(i_2, i_1) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -S_\lambda(i_l - 1, i_l) \\ -S_\lambda(i_l, i_1) & \dots & -S_\lambda(i_l, i_l - 1) & 1 - S_\lambda(i_l, i_l) \end{pmatrix}$$

$$et\ Y_I = \begin{pmatrix} \widehat{f}(x_{i_1}) - y_{i_1} \\ \vdots \\ \widehat{f}(x_{i_1}) - y_{i_l} \end{pmatrix}.$$

On peut donc généraliser à des cas où $K \neq n$ *avec le même raisonnement si la matrice* $I_l - S_\lambda(I)$ *reste inversible.*

Dans nos travaux, la validation croisée avec un nombre de "folds" bien choisi, a montrée une plus grande robustesse sur les données climatiques considérées, en comparaison à d'autres critères tels que AIC BIC. Il fallait néanmoins s'assurer que les "folds" étaient suffisamment indépendants les uns des autres. Pour se faire, nous avons choisi de tirer parmi des années entières, ce qui s'est avéré suffisant. En effet, une autre pratique pour les séries temporelles, est d'enlever les parties de l'échantillon d'apprentissage adjacentes au "fold" à estimer.

Chapitre 3

Étude de l'espérance d'une variable climatique au pas de temps journalier (déformation du cycle annuel)

Sommaire

Dans ce chapitre, nous mettrons à profit les méthodes de lissages splines et de validation croisée pour revisiter la notion de normale. La première partie est issue d'un article publié [109] et concerne la meilleure prédiction d'une normale pour l'année suivante. Elle se concentre sur l'étude de la température moyenne journalière, mais peut être appliquée à d'autres variables climatiques. Il s'agit donc d'obtenir la meilleure estimation de l'espérance d'une variable climatique dans le futur proche au vu des données actuelles.

La deuxième partie concerne l'estimation de normales sur une période observée (i.e. à l'intérieur de l'échantillon d'apprentissage). Dans notre cas, nous analyserons une période future dans le monde du modèle, afin de décrire plus finement les changements de saisonnalité. Ce travail se place donc dans le cadre de l'analyse du changement climatique (mis en opposition, ici, au cadre prédictif précédent). La question est donc d'obtenir la meilleure estimation de l'espérance de la variable climatique à l'intérieur de l'échantillon. Elle nous permettra notamment d'obtenir des moyens d'analyse du réchauffement saisonnier à travers une zone de l'hémisphère nord (centrée sur l'Europe).

Cette articulation correspond à l'utilisation habituelle des normales ; d'un côté nous avons la description de l'évolution de la valeur moyenne sur une période historique et d'un autre côté, la valeur moyenne espérée pour l'année suivante. Ces deux utilisations étaient issues du même diagnostic au cours du $19^{ème}$ siècle et de la première moitié du $20^{ème}$ siècle. En effet, le climat pouvant être supposé stationnaire au cours de ces périodes, la valeur moyenne à l'intérieur de l'échantillon est représentative de sa valeur en dehors de celui-ci. Le changement climatique impacte clairement ces deux aspects à l'échelle journalière. Plus précisément, la description d'un cycle annuel moyen admet, à présent, une évolution séculaire. Cela remet en question le calcul des cycles moyens pour chaque année mais surtout, la prévision de ceux-ci au cours des années suivantes.

3.1 Estimating daily climatological normals in a changing climate

Alix Rigal, Jean-Marc Azaïs, Aurélien Ribes

Article accepté par Climate Dynamics

Résumé

Les normales climatiques sont des références couramment utilisées pour la description et la caractérisation d'une situation météorologique donnée. Les normes de l'Organisation Météorologique Mondiale (OMM) recommandent d'estimer les normales comme une moyenne sur une période de 30 années. Cette approche peut mener, dans un climat qui change, à des normales fortement biaisées. Ici, nous proposons une nouvelle méthode avec laquelle nous pouvons estimer des normales climatiques au pas de temps quotidien en climat non-stationnaire. Notre cadre statistique s'appuie sur l'hypothèse que la réponse au changement climatique est régulière en fonction du temps, et que l'on a une décomposition de la réponse inspirée par une hypothèse du type "pattern scaling". L'estimation est effectuée en utilisant des techniques de lissage spline et en examinant soigneusement la sélection du paramètre de lissage. Cette nouvelle méthode est comparée, dans le sens prédictif et en mode modèle parfait, aux précédentes alternatives telles que les normales OMM standards (qu'elles soient remises toutes les décennies ou tout les ans) des moyennes sur des périodes plus courtes, ou encore du hinge fit. Nos résultats montrent que notre technique surpasse toutes les alternatives considérées. Ils confirment que les techniques précédentes sont substantiellement biaisées. Typiquement les biais vont de quelques dixièmes de degrés à plus de 1 degré à la fin du siècle, contrairement à celle que nous proposons. Nous soutenons que de telles normales "corrigées" pourraient être très utiles pour le suivi climatique, et que les services météorologiques pourraient considérer l'usage de deux types de normales (i.e. stationnaires et non-stationnaires) en fonction du but recherché.

Abstract

Climatological normals are widely used baselines for the description and the characterization of a given meteorological situation. The World Meteorological Organization (WMO) standard recommends estimating climatological normals as the average of observations over a 30-year period. This approach may lead to strongly biased normals in a changing climate. Here we propose a new method with which to estimate daily climatological normals in a non-stationary climate. Our statistical framework relies on the assumption that the response to climate change is smooth over time, and on a decomposition of the response inspired by the pattern scaling assumption. Estimation is carried out using smoothing splines techniques, with a careful examination of the selection of smoothing parameters. The new method is compared, in a predictive sense and in a perfect model framework, to previously proposed alternatives such as the WMO standard (reset either on a decadal or annual basis), averages over shorter periods, and hinge fits. Results show that our technique outperforms all alternatives considered. They confirm that previously proposed techniques are substantially biased. Biases are typically as large as a few tenth to more than 1 degree by the end of the century, while our method is not. We argue that such "climate change corrected" normals might be very useful for climate monitoring, and that weather services could consider using two different sets of normals (i.e. both stationary and non-stationary) for different purposes.

III.1.1 Introduction

Climatological normals are widely used baselines which describe and characterize a given meteorological situation. On the news, weather forecasts commonly refer to normals in order to compare a weather or seasonal forecast to its expectation. Retrospective climate monitoring also typically involves such a comparison. Climate normals are primarily meant to describe the mean seasonal cycle in standard meteorological variables such as temperature or precipitation. In the most common estimation techniques, normals are assumed to be stationary, i.e. the drift related to anthropogenic climate change is neglected, potentially leading to inaccurate or biased estimates. One issue with this approach is that, as pointed out by [9] : "climate normals are calculated retrospectively, but are often utilized prospectively". For instance, when they are compared to weather forecasts, it is assumed that normals provide an estimation of the

expected weather to date. Neglecting on-going warming can prevent this.

For example, the current recommendation of the World Meteorological Organization (WMO) for the calculation of climatological normals, known as the Climatological Standard Normals, is to compute an average over a 30-year period [138, 16]
(ref to `http://www.wmo.int/pages/prog/wcp/wcdmp/GCDS_1.php`). These normals are supposed to be updated every 30 years, with the current reference period being 1961-1990. Following these recommendations, the next generation of climatological normals would be available in 2021, based on the 1991-2020 average.

Several studies pointed out the limitation of such normals and their inaccuracy in a non-stationary climate (e.g [113, 85]). WMO itself advocated for a more frequent revision of climate normals, through updates every 10 years but still averaging over a 30-year period [139]. This change was intended to reduce the bias, with a careful discussion of pros and cons in order to define a dual standard for normals. Other authors, e.g. [93],[129],[130], proposed alternative methods for deriving climatological normals and assessed these methods across the US. Optimal Climatological Normals (OCNs) are averages calculated over periods shorter than 30-years. The length of the averaging period is then selected to maximize the accuracy of the estimation – the authors typically considered 15-year means for temperatures across the US. Hinge fits are break-point statistical models where the climatological mean is expected to be constant before a given date, and linearly growing after that date. Authors cited above suggest that 1975 is a good choice for the break point over the US. Some national meteorological services already use these alternative estimation techniques operationally (ref to https ://www.ncdc.noaa.gov/normalsPDFaccess/).

Another important feature of climatological normals is their time-resolution. Most normals are calculated on a monthly time-scale. However, for specific applications, the estimation of daily normals is required [10]. A few techniques have been proposed and /or are routinely used to translate monthly into daily values [8, 12, 11]. Another option is to estimate daily normals directly from raw data, assuming some type of regularity in the seasonal variations (i.e. normals do not vary much from one day to the next). Doing this in a non-stationary context will require smoothness both in the seasonal cycle and the climate change components. This is the method employed in this manuscript.

In this paper, we assess the accuracy of previously proposed techniques for the estimating of climatological normals. We outline their limitations if applied in the course of the 21st century. We then introduce a new approach for the overcoming of these issues. With this approach, the drift related to climate change on the seasonal component is estimated, leading to daily estimates. Also, the methodology can be applied to homogenized observations. All evaluations are made in a predictive sense, i.e. assessing whether normals calculated in the (recent) past provide a reliable estimation of current to near-future climates.

The manuscript is organized as follows. After presenting the dataset in Section III.1.2 we elaborate on the methods used to estimate climatological normals, then introduce our new method. The predictive skills of the various techniques considered are assessed and discussed in Section III.1.4. This is followed by a discussion along with some concluding remarks in the last section.

III.1.2 Data & existing methods

III.1.2.1 Data

In order to assess the accuracy of various techniques for estimating climatological normals during the 21st century – a period over which observations are not available – we use series of daily and annual mean temperatures. These are simulated by an ensemble of climate models from the Coupled Model Intercomparison Project Phase 5 (CMIP5) as realistic realizations of future observations. Estimation techniques are therefore compared in a perfect model framework (see more details in Section III.1.4).

More specifically, we focus on four locations which are meant to be representative of a wide range of climates : Bengaluru (India) in the tropics, Alert (Canadian Arctic Archipelago) in high-latitude, Paris (France) and San Francisco (California, USA) in mid-latitude regions. Twenty one CMIP5 models were selected for the daily mean temperature and sixty for the annual mean temperature (see Appendix A for a detailed list of models).

The considered time-series cover a period of 238 years from 1862 to 2099. They consist of the concatenation of two types of experiments :

- historical runs (driven by observed radiative forcings) covering the period 1862-2005,
- RCP8.5 scenario (Representative Concentration Pathways 8.5, corresponding to a high increase in greenhouse gas emissions during the 21st) simulations covering the remaining of the 21^{St} century (2006-2100).

The choice of a RCP8.5 scenario involves a strong climate change signal in the coming decades, but results obtained with this scenario are expected to hold at least qualitatively with more moderated alternatives.

It must be noted that for daily calculations, all the 29^{Th} February were removed to facilitate processing. Also, extensions to other climate variables, such as precipitation, are beyond the scope of this paper.

III.1.2.2 Previously introduced methods considered within this study

Here we review methods proposed by various authors in order to estimate climatological normals. Some of these techniques have been introduced in order to cope with climate change, and/or build upon the standard WMO recommendation. First we explain how these methods can be used to estimate annual normals, then we discuss how they can be extended to the daily timescale. This list of techniques is not meant to be exhaustive, but instead representative of what has been proposed in the literature.

- **III.1.2.2.1 WMO standard**

  The WMO recommendation is to calculate climatological normals as a simple average over a 30-year period composed of 3 full decades :

$$WMO(D+k) = \frac{1}{30} \sum_{i=D-29}^{D} T_i, \tag{3.1}$$

  where $D+k$ is the current year, D is the current decade (e.g 2010), $k \in [\![1, 10]\!]$ denotes the year within the decade, T_i is the mean temperature (or any other meteorological variable) of year i. This calculation is updated every 10 years which means that, after a

decade is completed, the estimated normals are valid and can be used for the subsequent 10 years (as denoted by $D + k$ in (3.1)).

- **III.1.2.2.2 WMO reset**

As a first very simple alternative, the same calculation can be made and updated every year (instead of every decade), leading to

$$WMO(y) = \frac{1}{30} \sum_{i=y-30}^{y-1} T_i, \tag{3.2}$$

where y is the current year. This will be referred to as WMO reset in the following, and is expected to be less biased than WMO in a changing climate thanks to the more frequent update.

- **III.1.2.2.3 Optimal Climate Normals (OCN)**

[62, 129, 130] argued that averaging over a 30-year period was non-optimal (too long) in a climate change context, and suggested tuning the length of the averaging period to improve the accuracy of the estimate. They suggested that averaging over the most recent 15 years was a good compromise for temperature normals. As follows, OCN therefore designates a 15-year average :

$$OCN(D + k) = \frac{1}{15} \sum_{i=D-14}^{D} T_i, \tag{3.3}$$

with $k \in [\![1, 10]\!]$. As for WMO, this average can be updated every 10 years (as assumed in the following), or every year. In the following, this 15-year average will be used as a benchmark for other operational normals using the mean of a different number of years.

- **III.1.2.2.4 Hinge fit**

In order to account for non-stationary climates, other authors proposed the use of a statistical model allowing for a trend in the estimation of climate normals. Among these, the most popular technique is the hinge fit [93, 130]. This is a simple break-point model

where the normals are assumed to be constant (i.e. non time-dependent) up to a given date, then linearly moving with time. The date of the break-point needs to be selected carefully – [93], [130] suggested that 1975 was an appropriate choice for the continental US and this is the value used in this paper.

$$Hinge(D+k) = \beta_0 + \beta_1 I_{1975}(D+k), \tag{3.4}$$

where $I(x) = 0$ if $x \leq 1975$ and $I(x) = x - 1975$ if $x > 1975$. The coefficients β_0 and β_1 are estimated from the full observational record available up to year D (i.e. not restricted to a 30-year period) using simple linear regression. Again this type of estimate could be updated each decade or year.

- **III.1.2.2.5 Hinge fit reset**

The same calculation can be made and updated every year instead of every decade, leading to

$$Hinge(y) = \beta_0 + \beta_1 I_{1975}(y), \tag{3.5}$$

The coefficients β_0 and β_1 are estimated from the full observational record available up to year $y-1$ using simple linear regression. This will be referred to as Hinge fit reset in the following.

III.1.2.2.6 From annual to daily normals

All of the techniques listed above provide annual or monthly normals ; an additional procedure has to be used to derive daily normals. Such a procedure was first introduced by [8], consisting of an expansion in a Fourier basis. We use a slightly different and simpler technique. Other approaches might have been used to derive daily normals ; an in-depth comparison of such methods goes beyond this study. Our technique is described below using the WMO estimate as an example, but it can be applied to any other estimator, including the OCN and Hinge methods introduced above.

Firstly, we compute normals for each single day within a year, i.e.

$$WMO_{raw}(D+k,d) = \frac{1}{30}\sum_{i=D-29}^{D} T_{i,d}, \tag{3.6}$$

where $d \in [\![1, 365]\!]$ represents the day, while other notations are consistent with (3.1). These daily values are then fitted onto the thirteen first elements of the Fourier basis. Equivalently, we estimate the linear coefficients α_i, β_i involved in the statistical model

$$WMO_{raw}(D+k,d) = \alpha_0 + \sum_{k=1}^{6}\left(\alpha_k \cos\left(\frac{2k\pi}{365}d\right) + \beta_k \sin\left(\frac{2k\pi}{365}d\right)\right) + \varepsilon_d. \tag{3.7}$$

Finally the estimated daily normals $\mathrm{WMO}_{\mathrm{day}}$ for year $D+k$ and day d are

$$WMO_{\mathrm{day}}(D+k,d) = \widehat{\alpha}_0 + \sum_{k=1}^{6}\left(\widehat{\alpha}_k \cos\left(\frac{2k\pi}{365}d\right) + \widehat{\beta}_k \sin\left(\frac{2k\pi}{365}d\right)\right), \tag{3.8}$$

where $\widehat{\alpha}_i, \widehat{\beta}_i$ are the estimated regression coefficients. Through projection onto a Fourier basis, this technique ensures regularity in the estimated annual cycle.

III.1.3 New Method

All methods described above could be criticized for a certain lack of flexibility (e.g. [84]). Indeed, Climate is either assumed to be stationary locally (computing averages) or moving linearly over time, with the linearity holding over a relatively long period of time, from 1975 onward (hinge fit). In this section, we introduce an alternative method for computing climatological normals, which is somewhat more flexible for it is based on spline smoothing. Obviously, the increase in flexibility is at the cost of an increase in the variance of the estimator - this will be discussed through the investigation of the overall performance of our approach in subsequent sections.

III.1.3.1 Statistical framework

The general statistical model considered is inspired by and adapted from [15]. Let $T_{y,d}$ be the mean (i.e. statistical expectation of) temperature of day d in year y. Our statistical model assumes that the following decomposition holds :

$$T_{d,y} = f(d) + g(y)h(d) + \varepsilon_{d,y}, \qquad d \in [\![1, 365]\!], y \in [\![1, n]\!], \tag{3.9}$$

where :

- $f(), g(), h()$ are smooth functions ($f(d), g(y), h(d)$ being their trace on integer values),
- $f(), h()$ are, additionally, periodic functions with period 365,
- ε is assumed to be Gaussian white noise with unknown variance σ^2.

In addition we impose the constraints $\sum_{y=1}^{n} g(y) = 0$ and $\sum_{d=1}^{365} h(d) = 1$ in order to ensure model identifiability (i.e. to avoid any possible confusion between the terms f and gh). Note that another system of constraints is possible in order to facilitate interpretation (see Appendix B).

This statistical model can be interpreted as follows. $f(d)$ represents a stationary seasonal cycle, which would be observed if the climate was stationary and the effect of climate change is described by the term $g(y)h(d)$. The key assumption is that this climate change response can be factorized into one component which describes how the shape of a seasonal cycle changes, $h(d)$, and another one which describes the variation of the magnitude of this change with time, in the long-term, $g(y)$.

This type of decomposition is an adaptation of the *pattern scaling* assumption [96, 120, 46] in a slightly different setup. Under *pattern scaling*, it is assumed that the spatial distribution of climate change does not vary with time – only the amplitude of the change does. It is thus possible to decompose climate change as the product between one spatial function, and one temporal function. In the present paper, the spatial component is replaced by the seasonal cycle. In both cases, the assumption can be thought of as a Taylor approximation of order one,

which is valid as long as the change is small enough. This factorization assumption is obviously one of consequence but has already proven its descriptive capabilities on hourly surface air temperature observations [125]. Its primary interest comes from the induced reduction in the model's complexity : estimating two univariate functions g and h is much easier than estimating a bivariate function (say $c(y, d)$). Its introduction therefore allows us to better constrain the estimation of the climate change component. If this assumption were invalid, the predictive performance of our method would be reduced, in particular if compared to methods not relying on a similar assumption. Lastly, model (4.1) proved a very good capability across the entire time series considered and it can be at least partially validated by examining goodness-of-fit to model (4.1), as discussed below.

An illustration of this model and the typical outputs it can produce, is shown in Figure 3.1. Next we will discuss how estimating the unknown functions f, g, h within this model. Goodness-of-fit of this model is also discussed in Section III.1.3.1.

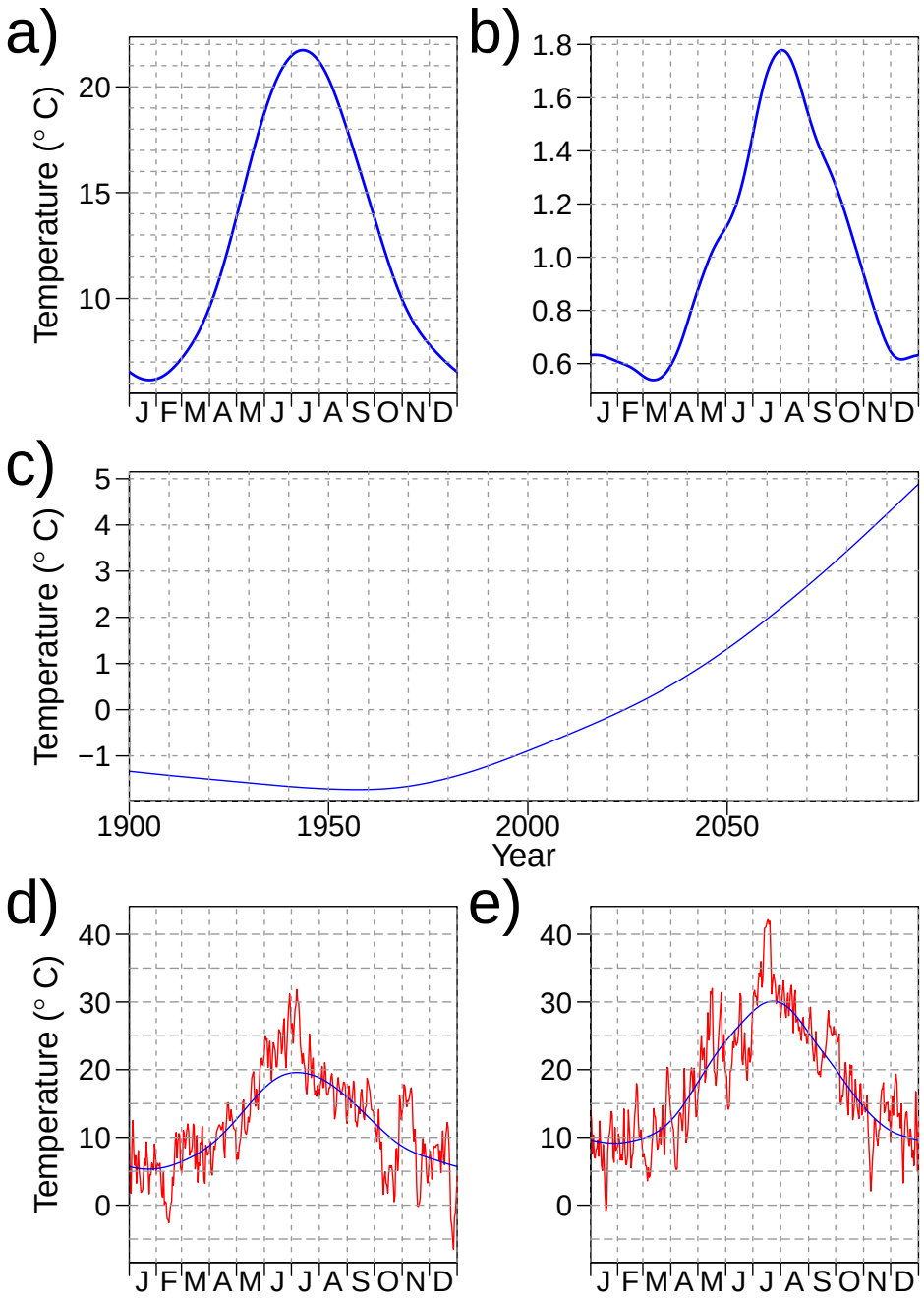


FIGURE 3.1 – Decomposition of a time series (Paris) by the spline model (4.1). a) represents the reference seasonal cycle f with *df=11* (see section 3.3), b) illustrates the seasonal drift h with *df=10*, and c) represents the annual trend g with *df=10*. The plots d) and e) show the estimation of the annual cycle in 1900 and 2100 respectively. Raw data are shown in red, while the fit of model (4.1) is in blue.

Estimation Algorithm

Our estimation procedure is a sequential, two-step procedure. Firstly, $g()$ is estimated using annual mean data only. Secondly, $f()$ and $h()$ are estimated assuming that $g()$ is known.

Both steps involve smoothing with cubic splines.
For instance, denoting $T_{.y} = \frac{1}{365}\sum_{d=1}^{365} T_{d,y}$ the annual mean temperature of year y, the smoothing splines estimate $\widehat{g}()$ of $g()$ can be defined as

$$\widehat{g}() = \underset{s()}{\operatorname{argmin}} \sum_{i=1}^{n} \left(T_{.y_i} - s(y_i)\right)^2 + \lambda \int_{y_1}^{y_n} \left(s''(x)\right)^2 dx, \tag{3.10}$$

where the minimum is taken over all possible function $s()$ belonging to the associated Sobolev space. A spline estimate thus performs a trade-off between closeness to input data (here $T_{.y}$), and roughness (last term in the right-hand side). λ is a regularization parameter determining the level of smoothness. The selection of λ is a common but difficult problem which is addressed in detail in Subsection III.1.3.1. Remarkably, the solutions of (4.2) are known in closed forms and can be computed easily.

Furthermore, we attempt to provide a calculation which meets operational constraints, and which is thus computationally not too expensive so as to apply it to multiple grid points. For this reason we implemented the sequential algorithm described below.

Algorithm

---

1 **Estimation of $g()$:**

Calculate the annual means $T_{.y}$. From the $T_{.y}$ time-series, compute the smoothing spline estimate $\widehat{g}()$ of $g()$, with a given df_g (smoothing parameter of g). Note that this estimate has to be centered subsequently in order to satisfy the identifiability constrains.

2 **Linear regression on $\widehat{g}(y)$:**

For each day $d \in [\![1, 365]\!]$, the time-series $T_{y,d}$ is linearly regressed onto $\widehat{g}(y)$, i.e. we estimate the coefficients α_d, β_d involved in :

$$T_{d,y} = \alpha_d + \widehat{g}(y)\beta_d + \varepsilon_{y,d}. \tag{3.11}$$

Thanks to orthogonality, $\widehat{\alpha}_d = T_{d,.}$, where $T_{d.} = \frac{1}{n}\sum_{y=1}^{n} T_{y,d}$, and $\widehat{\beta}_d = \dfrac{\sum_{y=1}^{n} \widehat{g}(y) T_{d,y}}{\sum_{y=1}^{n} \widehat{g}(y)^2}$.

3 **Estimation of $f()$ and $h()$:**

From the series $\widehat{\alpha}_d$ and $\widehat{\beta}_d$, respectively, we calculate the estimates $\widehat{f}()$ and $\widehat{h}()$, as periodic cubic smoothing splines estimates, with given df_f and df_h (smoothing parameters of f and h).

As it is sequential and based on an orthogonal design in the regression step, this algorithm is very rapid. A more sophisticated, iterated version of the algorithm has also been studied, and is presented in Appendix C. This variant showed no real improvement however and was thus dismissed.

Predictions based on the model (4.1) can be derived by extrapolating the estimated spline $\widehat{g}()$ to the year in question. Note that, as natural splines are used to estimate g (i.e second derivatives are null at the terminating points), this extrapolation is linear. The R-scripts used to carry out estimation of model (4.1) are available online via the CNRM-GAME website (URL : `http://www.umr-cnrm.fr/spip.php?article1064`).

Selecting degrees of freedom

The selection of the smoothing parameters λ (there is one parameter for each function $f(), g()$ and $h()$), will be discussed in terms of "equivalent degrees of freedom" (df), as in many spline papers. df is meant to be the equivalent of the number of parametric predictors involved in the estimation of the function. The smaller the df, the smoother the function estimate. Note that df is a complex one-to-one function of λ – but this correspondence will not be detailed further.

The determination of the different degrees of freedom (df) is performed using a variant of cross validation methods adapted to a prediction context. We use a multi-model ensemble

of transient simulations covering the 1850-2100 period (see Section III.1.2.1) as plausible realizations of the real world. From this dataset, we look for the value of *df* allowing the best prediction for the coming year (e.g. 2011) using data from previous years (e.g. 1850-2010).

This procedure is distinct from common cross-validation. Usual cross-validation would, in our case, consist of removing one or several years from the available observations (e.g. 1850-2017 if we are in 2018), and tuning the *df* coefficients to make the estimated normals as close as possible to the years removed. This procedure is then repeated by removing different years. If this type of cross-validation were used, then the *df* coefficients would be optimized to best estimate normals in the past – a period over which climate exhibits no or little change. Given that the non-stationary feature of climate is larger now than in the past, the best *df* for prediction might differ from the best *df* in the past – we checked that this was effectively the case.

Finally, the three coefficients involved, hereafter df_f, df_g, df_h, are estimated sequentially, instead of simultaneously. This makes the selection procedure computationally more affordable.

In each of the three cases, the observation is decomposed into a training sample and a testing sample. For various values of the number of degrees of freedom *df*, the considered function (f, g or h) is estimated on the training sample and then compared to the testing sample by measuring a Mean Square Error (MSE). Results are averaged over the available climate models. The *df* leading to the smallest MSE is then selected. In addition to the MSE value, we estimate its standard deviation which enables the computation of a plausible range of values for *df*, through the *one standard error rule* [53].

Degree of freedom of the reference cycle $f()$

$f()$ is meant to represent the mean annual cycle in a stationary climate. In order to select df_f, we took the periods 1900-1930 as a training sample, and 1931-1940 as a testing sample. This somewhat subjective choice was motivated by the fact that climate in the early 20^{th} century is almost stationary.

The selected df_f is typically between 10 and 20, depending on the location considered. Note that the signal-to-noise ratio is much higher for this stationary component $f()$ than for the remaining $g()$ and $h()$ functions, which explains why df_f is relatively large and well defined.

Degree of freedom of the annual trend $g()$

Unlike df_f, df_g depends on the decade considered. For a given decade D (for example 2001-2010), we use the data prior to D (i.e. the period 1862-2000) as a training sample, the decade itself being the testing sample. Again, we use the *one standard error rule* to assess a range of value for df_g.

Our selection procedure for df_g is illustrated in Figure 3.2a–b. Note that only annual mean values are used there. Focusing on the 2050 decade, the best value for df_g lies between 5 and 6 (panel a). Values smaller than 3 are clearly discarded, but the accuracy of the estimated normals is only slightly deteriorated if larger df_g are used (up to more than 15). Remarkably, the selected df_g is almost constant from 1990 to 2100 (panel b), with optimal values around 6. This applies to many other locations (not shown). Moreover, as the cross validation curve was very flat around its minimum, for all predictions made after 1990, we will use $df_g = 6$ in the following. Using such a constant value makes the algorithm easier to implement.

Degree of freedom of the delta cycle $h()$

Once df_f and df_g have been determined, estimates of $f()$ and $g()$ can be derived, and df_h is the only missing parameter to fit. In order to select df_h for a given year (2018 for example), we used *the past* (i.e. 1862–2017) as a training sample, then calculate the mean square error (MSE) over the next year (2018 in this case). Due to the strength of internal variability, we applied a smoothing over time. For each year, the selected df_h is the one minimising the smoothed MSE (see Figure 3.2c–d).

Our results suggest that df_h is the most sensitive (and therefore difficult to estimate) parameter in our statistical model. The selected values for df_h vary substantially both over

space and time. In the case of Paris (Figure 3.2d), df_h increases with time from near 1 (i.e. the minimum possible value, corresponding to no change in the annual cycle) to 10 in 2100. This corresponds to the signal-to-noise increase across the 21st century. In 1990, climate change was limited, and it is unclear which season experienced the greatest warming. It is thus safer to assume a flat response (i.e. the same degree of warming throughout the year). During subsequent decades, this changing signal (including change in the annual cycle) becomes clearer and greater flexibility in $h()$ becomes effective.

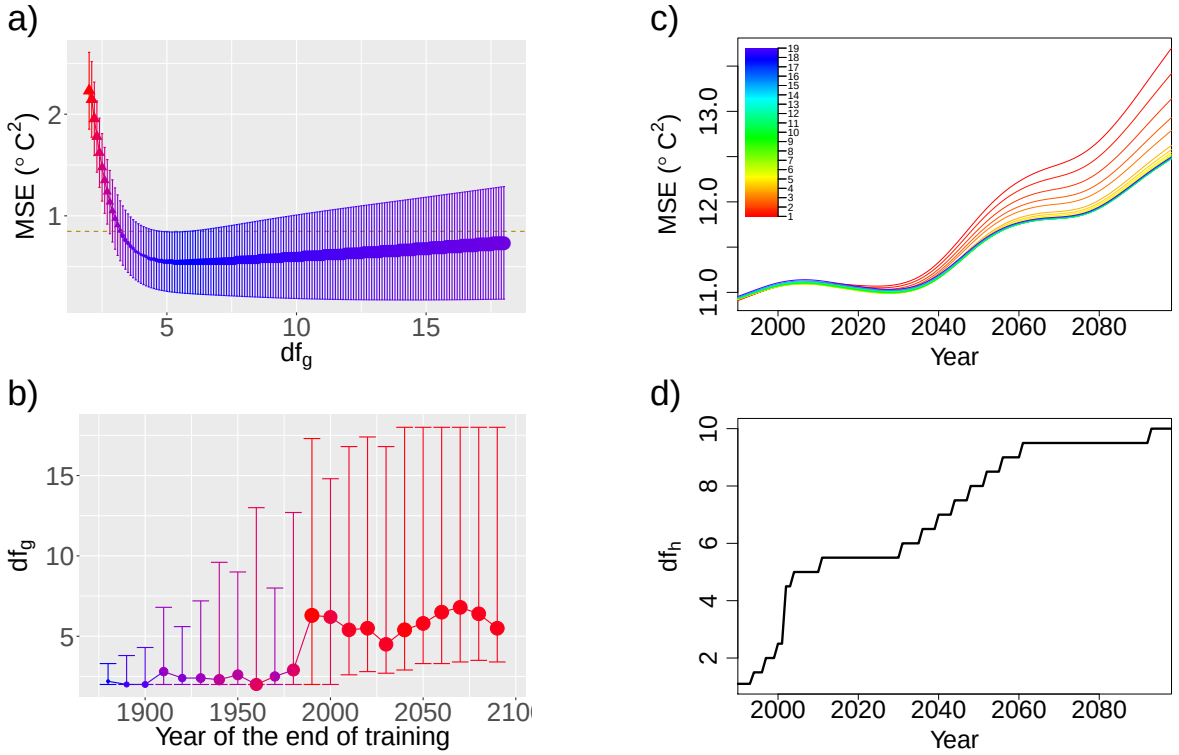


FIGURE 3.2 – **Selecting df_g and df_h. a)** error of annual mean temperature normals (points) for the decade 2050 (normals are estimated from 1862-2050 data, error is calculated over 2051-2060), and its standard deviation (bars), as a function of df_g. **b)** selected df_g (points), with the corresponding uncertainty according to the one standard error rule (bars), as a function of the predicted decade D. **c)** Mean square prediction error of daily normals as a function of time, for different values of df_h (df_g, df_f are given). **d)** Selected df_h as a function of the predicted year. All calculations in this figure are made for the Paris (France) grid-point.

Model goodness-of-fit

An important step in order to validate the use of our statistical model (4.1) and its underlying assumption is to assess the goodness-of-fit to this model. This can be done using climate model

data, and fitting the model across the entire period considered (1862–2100). Such diagnoses are shown in Figure 3.3. Consistent with Figure 3.1, these diagnoses apply to Paris and the CNRM-CM5 climate model ; they are representative of different locations and models.

Firstly, the determination coefficient $R^2 = 0.73$ is relatively high, and consistent with the internal climate variability. Residuals show no abnormal patterns : the Gaussian assumption is reasonably well-satisfied (3a), and they do not exhibit clear dependence on the fitted value (3b). Note that in the latter panel, the density of points depend strongly on the fitted values, thanks to the annual cycle and the fact that the climate is almost stationary over the first 100 years. For instance, the accumulation around $19°C$ is due to pre-industrial summer maxima.

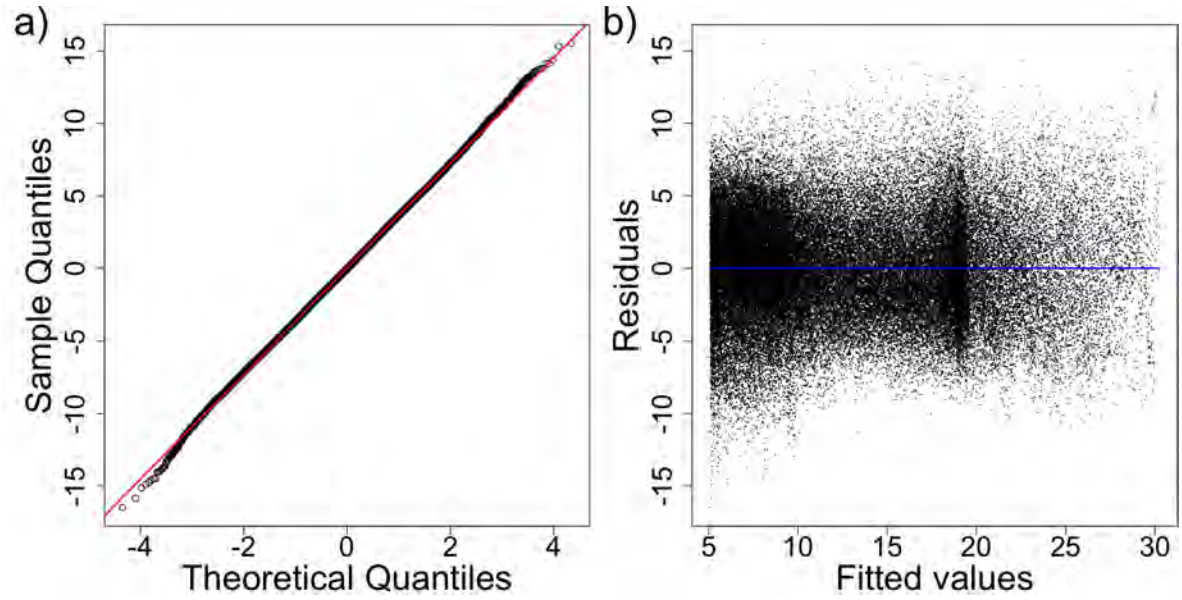


FIGURE 3.3 – Goodness-of-fit to our statistical model. **a)** normal QQ-plot of the residuals. **b)** residual vs fitted values plot.

III.1.4 Results

Scores on annual mean temperature

The results of the five methods introduced above are compared for annual mean temperatures, taken as pseudo observations, in Figure 3.4. The comparison is performed for 4 distinct locations, corresponding to different climates (mean temperature ranges from −32℃ to +27℃), amounts of warming (from 4℃ to 8℃ in 2100 under RCP8.5), and signal-to-noise ratio (inter-

nal variability being relatively smaller in the tropics). In addition, a smoothing spline of the entire time series is computed for comparing the bias of each method. The degree of freedom for each location is established by usual cross validation.

Globally, for all locations, the methods have almost the same performance until the late 20^{th} century (near 1990 or 2000, depending on the location). The hinge fit however seems to exhibit a larger variance after 1975 (see e.g. quick variations in Alert and Bengaluru). This is because very few points contribute significantly to fit the broken line's trend. This also applies to a lesser extent to OCN, given that the average is calculated on a smaller number of years than that of the WMO. The sampling margin of error is therefore larger.

During the early 21st century, methods based on averaging over past years (namely OCN, WMO, and WMO reset), are starting to depart from the reference, and show a negative bias. Hinge fit and our technique do a much better job and remain close to the reference.

Lastly, our method performs much better than any other in the second half of the 21st century. While this method remains continuously close to the reference, alternatives systematically underestimate the current state of the climate, by .5 to 1 degree.

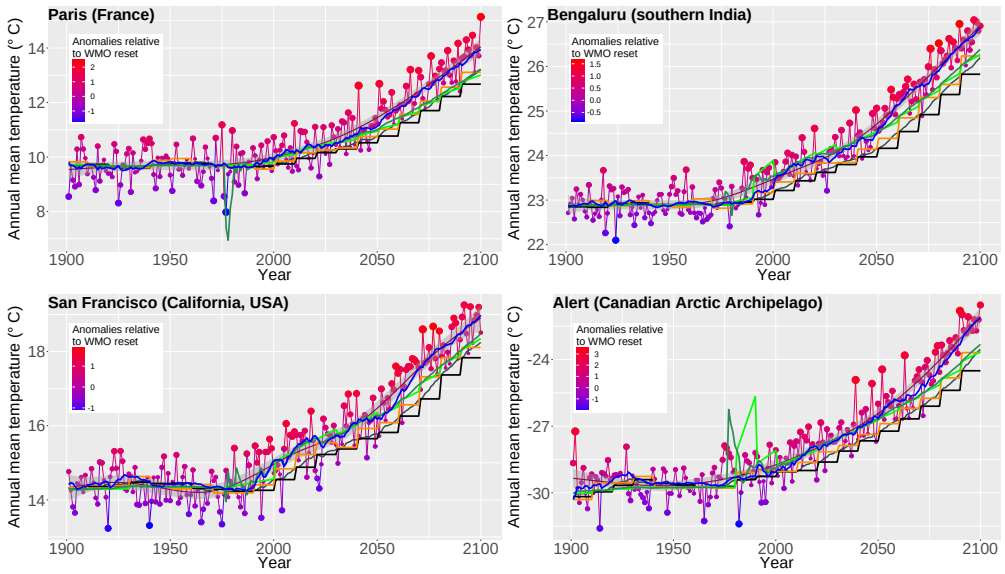


FIGURE 3.4 – Annual mean temperature and estimated normals. Temperature normals on an RCP8.5 scenario of the CNRM-CM5 model. Time-series of annual mean temperature (points) at four different locations (panels). Climatological normals are estimated using 6 different techniques : WMO standard (black line), WMO reset (grey), OCN (yellow), hinge fit (light green), hinge fit reset (dark green), and our method (blue). Normals for a given year (e.g. 2018) are estimated using data from previous years (e.g. 1900-2017). A smoothing spline of the entire time-series (1900-2100 ; purple line) can be considered as a reference. Anomalies of individual years are calculated with respect to the WMO reset, in order to further illustrate the bias related to this method. All calculations are based on one RCP8.5 simulation from CNRM-CM5.

Beyond the illustrative and qualitative assessment made in Figure 3.4, methods can be quantitatively compared using standard criterion such as MSE, bias and variance. Such a score-based comparison will be carried out in detail in the next section for daily normals. It is also appropriate on annual time-scales, and is illustrated in appendix. These quantitative results are consistent with those described above.

Scores on the daily timescale

Figure 3.5 compares the performance of the 5 considered methods, at the daily timescale, and for one grid point near San Francisco. Again, the estimation techniques are trained on all years prior to the one predicted, then the estimated normals are extrapolated to that year. Evaluation of the methods is based primarily on the mean square error (MSE)[30]. The latter

is also decomposed as the sum of the $bias^2$ and the variance.

Bias varies from 0 to more than 1℃ depending on the method and period of time. If all methods are nearly unbiased in the late 20^{th} century, only our approach remains unbiased throughout the 21st century. Alternatives exhibit negative bias as large as .5 to .8℃, except for the standard WMO approach for which the bias is even larger, near or beyond 1℃. Even though hinge and hinge reset lie close to our method until 2040, their bias are slightly larger, on average. Overall, in the 2000-2100 period, methods can be sorted with respect to their bias (increasing order) : our method, hinge reset, hinge, OCN, and WMO. These results are highly consistent with those obtained on annual mean temperature.

The variance of all estimation techniques are in fact very close to one another. Only the hinge and hinge reset estimators yield a slightly higher variance than others, especially near the beginning of the period. Our technique has the lowest variance on average over the entire period.

In terms of Mean Square Error (MSE), which is an aggregation of bias and variance and a very usual criterion, our technique performs much better than all proposed alternatives. OCN and WMO approaches are reasonably accurate near the beginning of the period, for instance before 2020, when climate change remains slight. They are penalized by their large bias subsequently. The two variants of the hinge technique suffer from their large variance at the beginning, then rank second from 2010 to 2020.

A few additional remarks can be made. Firstly, results found for other locations were qualitatively similar. In particular, they confirm that our method outperforms the proposed alternatives, and remains almost unbiased across the 21st century. Secondly, all methods reset on a decadal basis exhibit some degradation of their scores at the end of each decade (WMO, OCN and hinge). This is particularly pronounced in the bias of OCN and WMO. Thirdly, these daily results are probably mainly explained by a better fit of the secular trend (as opposed to the annual cycle). Whether or not our technique provides a better estimation of the annual cycle itself was not investigated. Lastly, it should be noted that hinge fit paradigm would benefit from breakpoint analysis to reduce its bias on the second half of the 21 st century. However, knowing in what manner it would achieve a better trade-off is out of the

scope of this paper and would require complementary analysis.

Overall, these results suggest that our method is more accurate than existing alternatives to estimating daily normals. This happens both in terms of bias and variance, which can be underlined. Furthermore, very low bias is revealed over the 21st century. This suggests that our technique has the appropriate level of flexibility to follow climate change, whilst not having too much variance. As our method exhibits almost no bias, potentially more sophisticated methods could improve on the variance (the bias is already minimal). This would probably lead to limited gain in terms of total MSE, as a large part of this variance is related to (irreducible) internal climate variability.

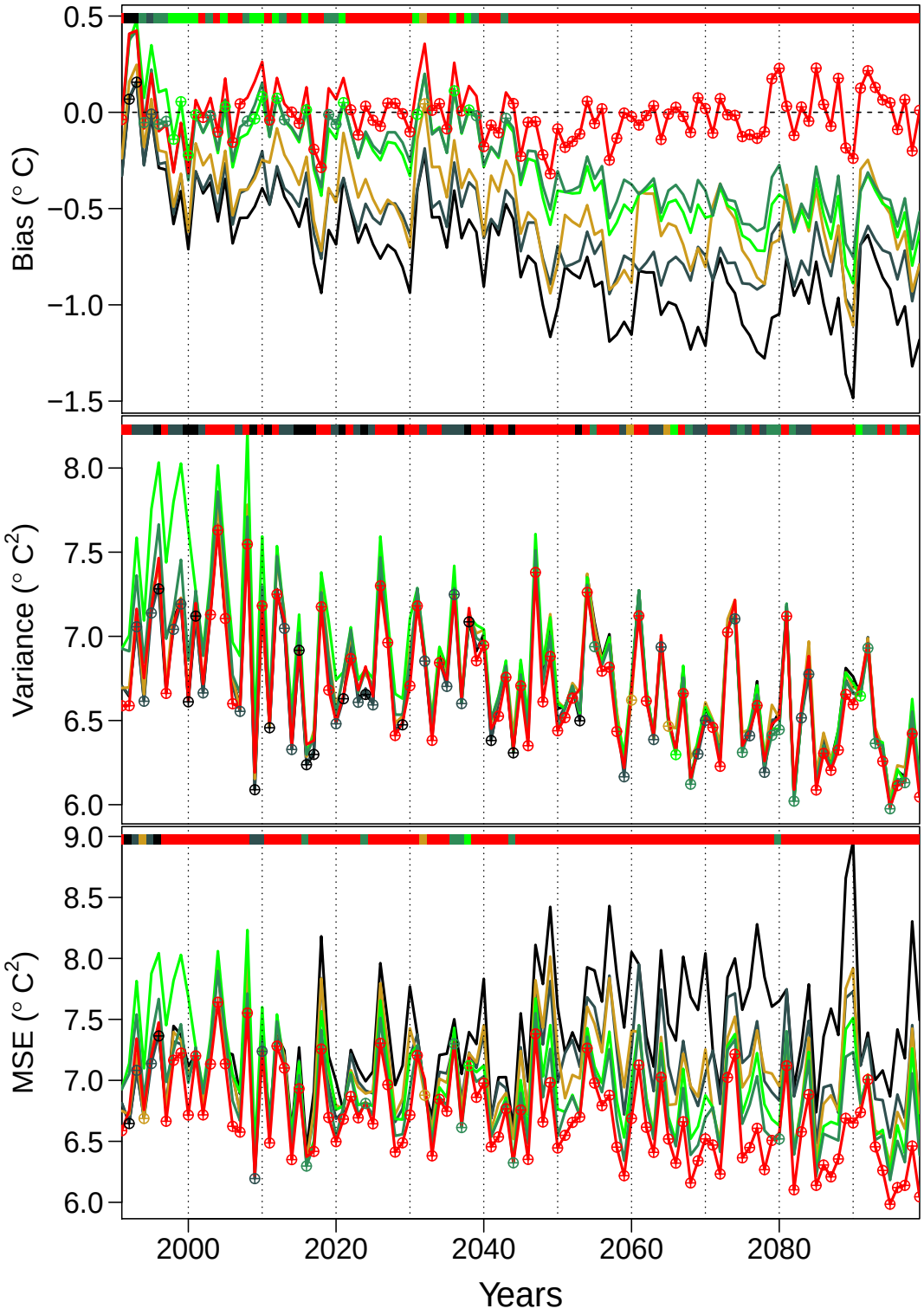


FIGURE 3.5 – Daily scores in San Francisco. Six techniques for estimating daily normals, namely WMO standard (black), WMO reset (grey), OCN (yellow), hinge fit (light green), hinge fit reset (dark green), and our method (red), are compared. Their evaluation is based on their bias (top), variance (middle), and mean square error (MSE, bottom). The year in the x-axis denotes the end of the training period ; prediction is made for the following year. The coloured line (top of each panels) indicates which method performs best, for a given criterion and a given year. Calculations are made for one grid-point near San Francisco, using an ensemble of RCP8.5 simulation from the CMIP5 archive.

III.1.5 Conclusion and discussion

In this paper, we introduce a new method for estimating daily climatological normals, describing the corresponding model. This technique relies on the assumption that the response to climate change is smooth over time and nears the pattern scaling assumption. All terms can by estimated using smoothing splines. The proposed estimation algorithm is very fast and this is due to a two-step (as opposed to simultaneous) procedure. The main challenge is the tuning of the smoothing parameters which is done using an extension of cross validation specifically designed for prediction. Once these smoothing parameters have been estimated or fixed, the method can be applied on a set of real homogenized observations.

Our method is compared to previously proposed alternatives in a predictive sense : methods are used to estimate climatological normals for the next, unobserved year. Their accuracy is compared on that basis, using an ensemble of RCP8.5 simulation from the CMIP5 ensemble in a perfect model framework.

Results show that our method is more accurate than all considered alternatives on the yearly timescale. The gap is particularly large across the second part of the 21^{St} century. Additionally, on the daily timescale, our method was also shown to provide the best results in terms of bias, variance, and therefore mean square error. These good properties can be partly attributed to the flexibility of the method, adjusted through the selection of smoothing parameters.

Our results thus suggest that the proposed method brings a strong improvement in the estimation of climatological normals accounting for climate change. Such revised – with respect to the WMO recommendation – normals could be used to address several questions. Unbiased normals could be particularly useful for climate monitoring, e.g. qualify if a year or season is warmer or colder than *really* expected. It could also be used to produce climate change corrected times-series. This would be relevant e.g. to compare how anomalous different years or periods are. As a typical illustration, one might wonder if an extreme event like the 2003 European Event remains unprecedented after correction for the climate change effect. Additionally, our method could be used to provide a refined description of on-going climate change with respect to the annual cycle, i.e. beyond the annual mean warming.

These attractive features do not mean that the standard way of computing climatological normals is now obsolete. Having a stationary reference such as the WMO standard is still very valuable, e.g. in order to highlight climate change. We suggest therefore that weather or climate services in charge of climate monitoring could compute two different sets of normals – a stationary reference and a climate change corrected set of normals – and use one or the other depending on the application considered. Updating the revised set of normals on a regular annual basis seems to be something required for the delivery of an estimation as accurate as possible.

Future work on the method described in this paper could include the estimation of uncertainties in the estimated normals. This would be very valuable, e.g. for assessing the uncertainties in climate change corrected time-series. Future work could also include a pre-computation of smoothing parameters for a large number of locations, in order to make the method even easier to implement. This tuning step remains the most difficult in our procedure and has to be re-examined carefully for different places. Lastly, the selection of smoothing parameters could be re-examined for different emission scenarios for the shape of the time response (and therefore the optimal value of smoothing parameters) greatly depends on the emission pathway.

III.1.6 Appendix

The analysis in this article has been performed using the statistical software `R`.

III.1.6.A Computational and simulation details

The 21 simulations used for daily mean temperature were :
ACCESS1-0, ACCESS1-3, CCSM4, CESM1-BGC, CMCC-CMS, CNRM-CM5, CSIRO-Mk3-6-0, CanESM2, GFDL-CM3, GFDL-ESM2G, GFDL-ESM2M, IPSL-CM5A-LR, IPSL-CM5A-MR, IPSL-CM5B-LR, MIROC-ESM-CHEM, MIROC-ESM, MPI-ESM-LR, MPI-ESM-MR, MRI-CGCM3, NorESM1-M, inmcm4

The simulations used for annual mean temperature were :
ACCS0 _ r1i1p1, ACCS3 _ r1i1p1, BCCl _ r1i1p1, BCCm _ r1i1p1, BNU _ r1i1p1, CCCMA _ r1i1p1, CCCMA _ r2i1p1, CCCMA _ r3i1p1, CCCMA _ r4i1p1, CCCMA _ r5i1p1, CNRM _ r10i1p1, CNRM _ r1i1p1, CNRM _ r2i1p1, CNRM _ r4i1p1, CNRM _ r6i1p1, CSIRO _ r10i1p1, CSIRO _ r1i1p1, CSIRO _ r2i1p1, CSIRO _ r3i1p1, CSIRO _ r4i1p1, CSIRO _ r5i1p1, CSIRO _ r6i1p1, CSIRO _ r7i1p1, CSIRO _ r8i1p1, CSIRO _ r9i1p1, GFDLc _ r1i1p1, GFDLg _ r1i1p1, GFDLm _ r1i1p1, GISSr _ r1i1p1, IAPg _ r1i1p1, IAPs _ r1i1p1, IAPs _ r2i1p1, IAPs _ r3i1p1, INGVc _ r1i1p1, INGVe _ r1i1p1, INGVs _ r1i1p1, INM _ r1i1p1, IPSLal _ r1i1p1, IPSLal _ r2i1p1, IPSLal _ r3i1p1, IPSLal _ r4i1p1, IPSLam _ r1i1p1, IPSLb _ r1i1p1, MIROC5 _ r1i1p1, MIROC5 _ r2i1p1, MIROC5 _ r3i1p1, MIROCc _ r1i1p1, MIROCe _ r1i1p1, MPIMl _ r1i1p1, MPIMl _ r2i1p1, MPIMl _ r3i1p1, MPIMm _ r1i1p1, MRI _ r1i1p1, NCARc _ r1i1p1, NCARc _ r2i1p1, NCARc _ r3i1p1, NCARc _ r4i1p1, NCARc _ r5i1p1, NCARc _ r6i1p1, NCARe _ r1i1p1

III.1.6.B Another system of constraints for model (4.1)

Once we have obtained the decomposition of model (4.1), it is possible to make it more interpretable. Let $\tilde{g} = g - g(1), \tilde{f} = f + g(1).h$. Then, the decomposition of model (1) can be rewritten as :

$$f(d) + g(y).h(d) = (f(d) + g(1).h(d)) + (g(y) - g(1)).h(d)$$
$$= \tilde{f}(d) + \tilde{g}(y).h(d)$$

Thus, $\tilde{f}$ represents the annual reference cycle of the first year of the considered period and $\tilde{g}$ quantifies the annual mean temperature evolution. Therefore the first value, $\tilde{g}(1)$, is zero.

III.1.6.C Alternating least squares

Addition of a few steps to the sequential algorithm permitting an iterative procedure :

4 **Re-estimation of $g()$:**

We now fix $\hat{f}, \hat{h}$ and estimate g once again, the goal of the procedure being minimization of the total sum of squares

i.e $RSS = \sum_{y,d}(T_{y,d} - \hat{f}_d - \hat{g}_y.\hat{h}_d)^2$.

For a fixed y, let us define :

$RSS_y = \sum_d((T_{y,d} - \hat{f}_d) - g_y.\hat{h}_d)^2 = \sum_d(\widetilde{T}_{y,d} - g_y.\hat{h}_d)^2$

where $\widetilde{T}_{y,d} = T_{y,d} - \hat{f}_d$

let $g_{0,y}$ the mean square estimator $g_{0,y} = \frac{\sum_{j=1}^{365} \hat{h}_d.T_{y,d}}{\sum_{j=1}^{365} \hat{h}_d^2}$

Also by the Pythagorean theorem :

$$\sum_{d=1}^{365}(\widetilde{T}_{y,d} - g_y.\hat{h}_d) = \sum_{d=1}^{365}(\widetilde{T}_{y,d} - g_{0,y}.\hat{h}_d + (g_{0,y} - g_y).\hat{h}_d)^2$$
$$= \sum_{d=1}^{365}(\widetilde{T}_{y,d} - g_{0,y}.\hat{h}_d)^2 + \sum_{d=1}^{365}((g_{0,y} - g_y).\hat{h}_d)^2$$
$$= \sum_{d=1}^{365}(\widetilde{T}_{y,d} - g_{0,y}.\hat{h}_d)^2 + (g_{0,y} - g_y)^2.\sum_{d=1}^{365}\hat{h}_d^2$$

Finally,

$$\begin{aligned} RSS &= \sum_{y=1}^{n} RSS_y \\ &= \sum_{d,y} (\widetilde{T}_{y,d} - g_{0,y}.\hat{h}_d)^2 + \sum_{y=1}^{n} (g_{0,y} - g_y)^2 . \sum_{d=1}^{365} \hat{h}_d^2 \end{aligned}$$

Then, we compute the smoothing spline estimate $\widehat{g}()$ of $g_{0,y}$, with the given df_g.

5 We iterate steps 3 and 4 to minimize sum of squares RSS.

III.1.6.D Annual scoring for normals

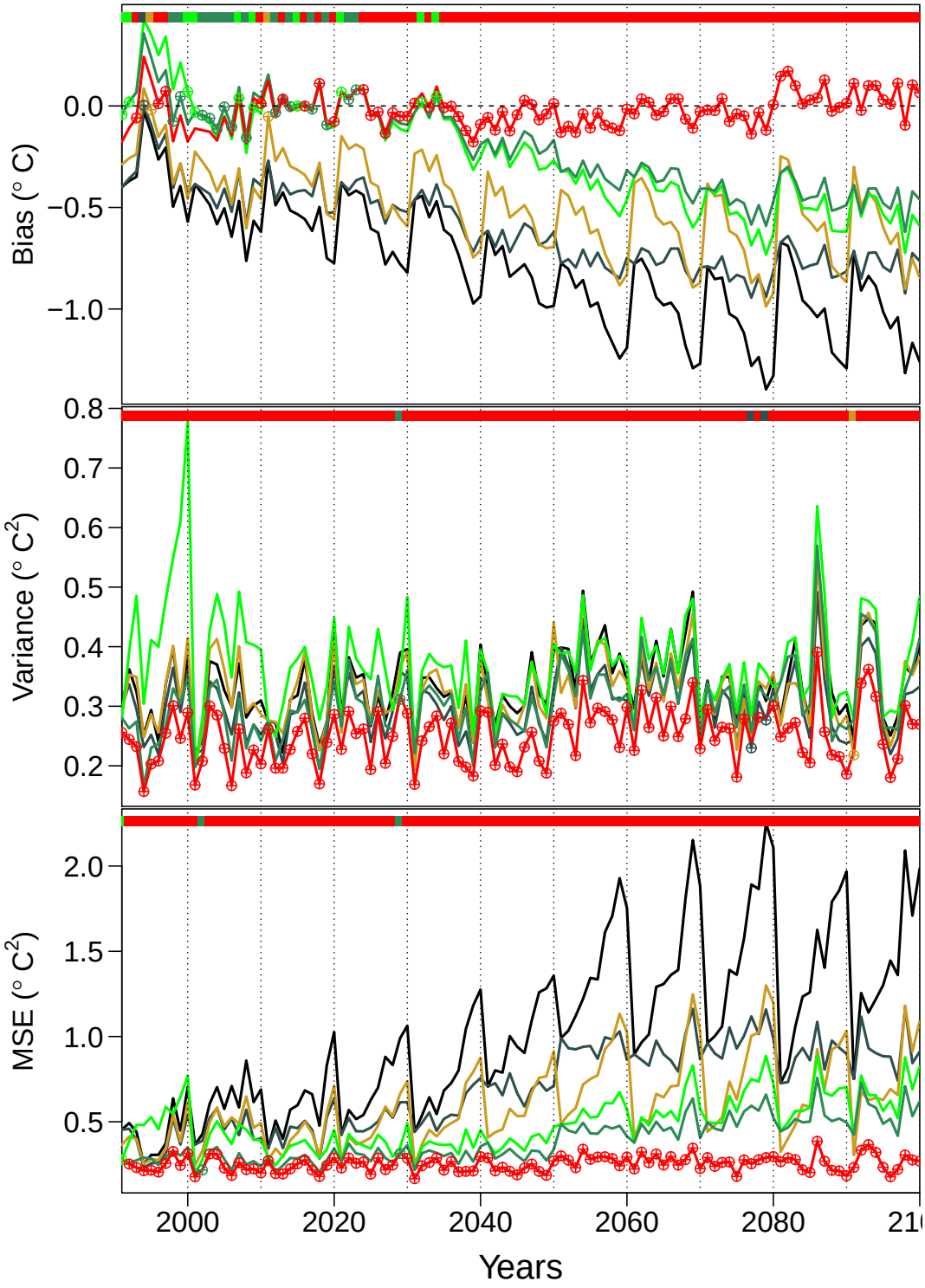


FIGURE 3.6 – Please see next page for description.

The three plots illustrate scoring on the yearly mean temperature at San Francisco, for each year normal prediction occurs on all CMIP5 models. The horizontal axis represents the end of the training period and for each method, prediction occurs the following year. The upper line shows, for each score, the winning method for predicting the next year. The different calculations are WMO (black), WMO reset (grey), OCN (yellow), hinge (light green), hinge fit reset (green) and model(4.1) (blue). The upper figure shows the evolution of the bias, the middle one represents the variance of the prediction and the bottom plot illustrates the evolution of the mean square prediction error (MSE).

Acknowledgements

The authors acknowledge Météo-France for supporting this study. They also wish to thank the climate modeling groups involved in CMIP5 for producing and sharing their simulations.

III.1.7 Supplementary materials

III.1.7.1 Does it still work on other RCP scenario ?

We could think that the performances of the new method are simply due to curve fitting of the secular warming, and it is true at the first order. However, we find a non-uniform delta cycle h (Figure 3.2), suggesting that changes in the annual cycle bring some contribution to the overall results.

Additionally, the outperformance of our method is robust when considering a RCP4.5 scenario where the future warming is not accelerating anymore (the warming trend is even reduced in the late 21st century in such a scenario) – see Figure 3.7.

Admittedly, the results on a RCP2.5 scenario would be less remarkable because the bias of other methods would be less pronounced. However, our method could still make a better bias-variance trade off because of its adaptability to multi-decadal variations. In addition, acceleration of warming seems to occur with a different timing and shape depending on the location (see for example Figure TS.12 of Assessment Report 5, working group 1). For instance, on a RCP2.6 scenario where temperature would stabilize, the Hinge fit would lack flexibility and overestimate normals.

Furthermore, as shown in Figure 5, Hinge yields bigger variance near its Hinge point, so less bias is sacrificed to the cost of bigger variance of the estimator.

FIGURE 3.7 – Please see next page for description.

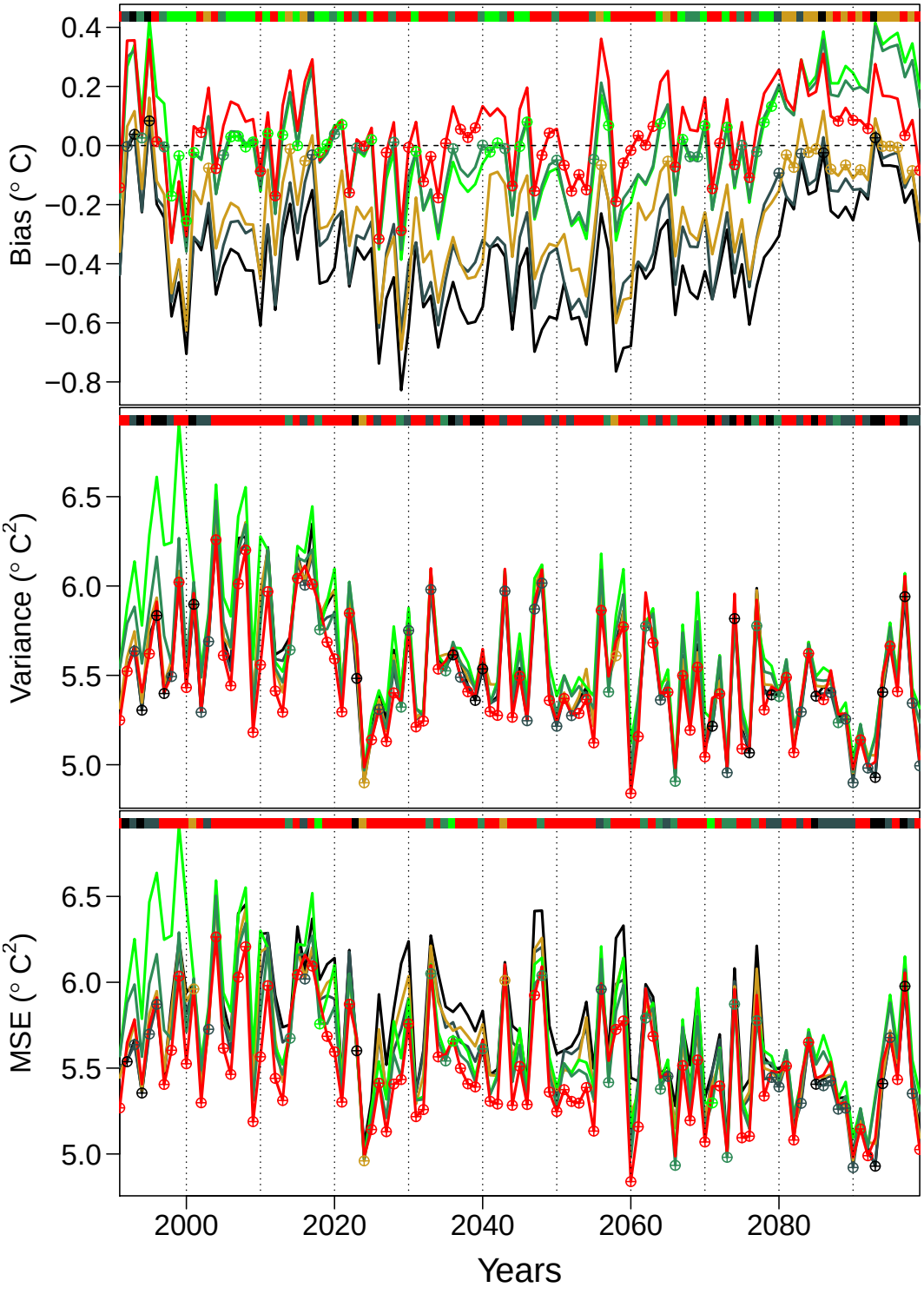


Figure 3.7 shows daily scores in San Francisco for a RCP 4.5 scenario. Six techniques for estimating daily normals are compared, namely WMO standard (black), WMO reset (grey), OCN (yellow), Hinge fit (light green), Hinge fit reset (dark green) and our method (red). Their evaluation is based on their bias (top), variance (middle) and mean square error (MSE, bottom). The year in the x-axis denotes the end of the training period ; prediction is made for the following year. The coloured line (top of each panels) indicates which method performs best, for a given criterion and a given year. Calculations are made for one grid-point near San Francisco, using an ensemble of RCP4.5 simulation from the CMIP5 archive.

III.1.7.2 Decomposition over the different considered locations

More examples of the estimations of daily temperature normals made by model 4.1 can be found below in the following figures (3.8,3.9, 3.10). The normals are calculated on our in-house model CNRM-CM5 at the different locations considered in the article. Although all localizations are taken in the Northern Hemisphere, annual cycles and their evolution are quite different.

Bengaluru situated in southern India's yields the flattest annual cycles. The results show features of observed climate change in this region for mean temperature [89] such as a not very seasonally marked climate change. Doing so, the annual cycle's shape, even on an RCP8.5 scenario, is nearly the same in 2099 and 1900.
Near Alert, in the Arctic Archipelago, we observed the biggest changes in temperature. As it is well-known the norther regions are the most impacted. Furthermore, warming was non-uniform across the year, as shown in figure 3.9. Summer warming was much weaker than in winter. This is very consistent with observed climate change in Canada [144]. Our estimations shows the disappearance of extreme cold days with mean temperature below $-20°C$ and the apparition of 5 months where temperature exceeds $0°C$ at the end of the century. California's seasonal change shows less warming in the end of the winter season and a stronger signal in June and September. This leads to longer summer and winter seasons in 2100.

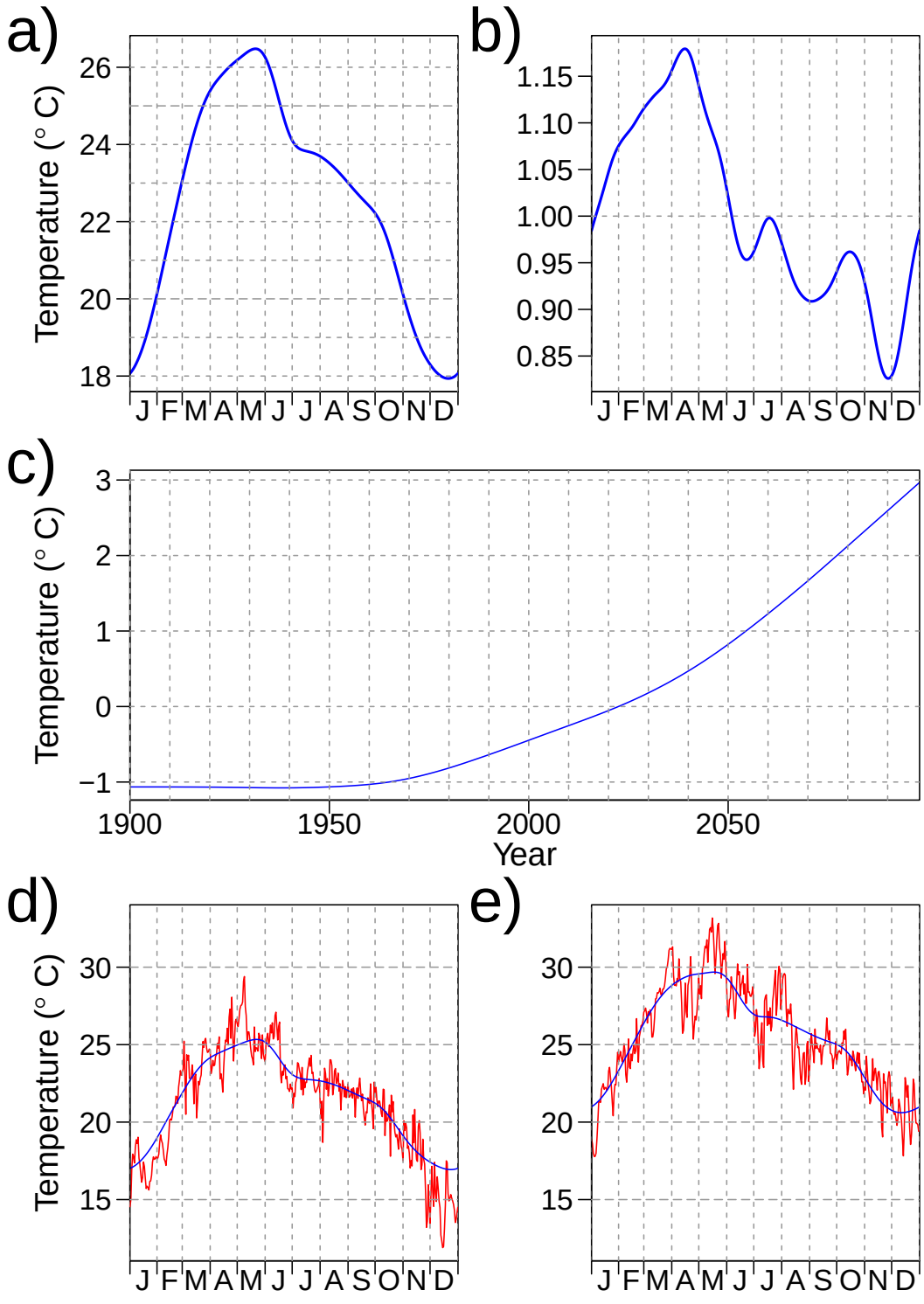


FIGURE 3.8 – Decomposition of a time series (Bengaluru) by the spline model (9) trained on the 1900-2099. a) represents the reference seasonal cycle f with *df=16.2*, b) illustrates the seasonal drift h with *df=12*, and c) represents the annual trend g with *df=6*. The plots d) and e) show the estimation of the annual cycle in 1900 and 2100 respectively. Raw data is shown in red, while the fit of model (9) is in blue.

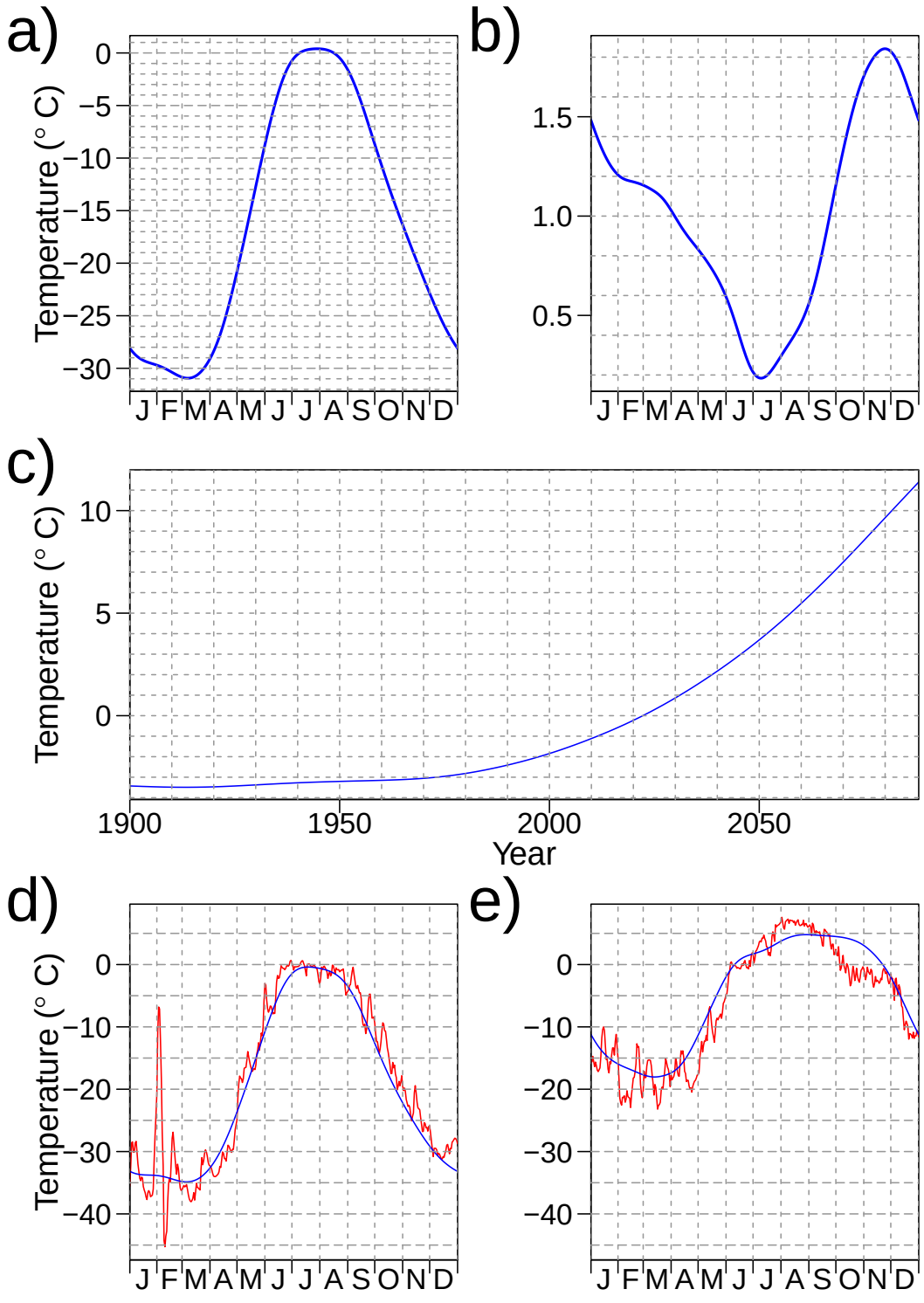


FIGURE 3.9 – Decomposition of a time series (Alert) by the spline model (9) trained on the 1900-2099. a) represents the reference seasonal cycle f with *df=13.2*, b) illustrates the seasonal drift h with *df=12*, and c) represents the annual trend g with *df=6*. The plots d) and e) show the estimation of the annual cycle in 1900 and 2100 respectively. Raw data are shown in red, while the fit of model (9) is in blue.

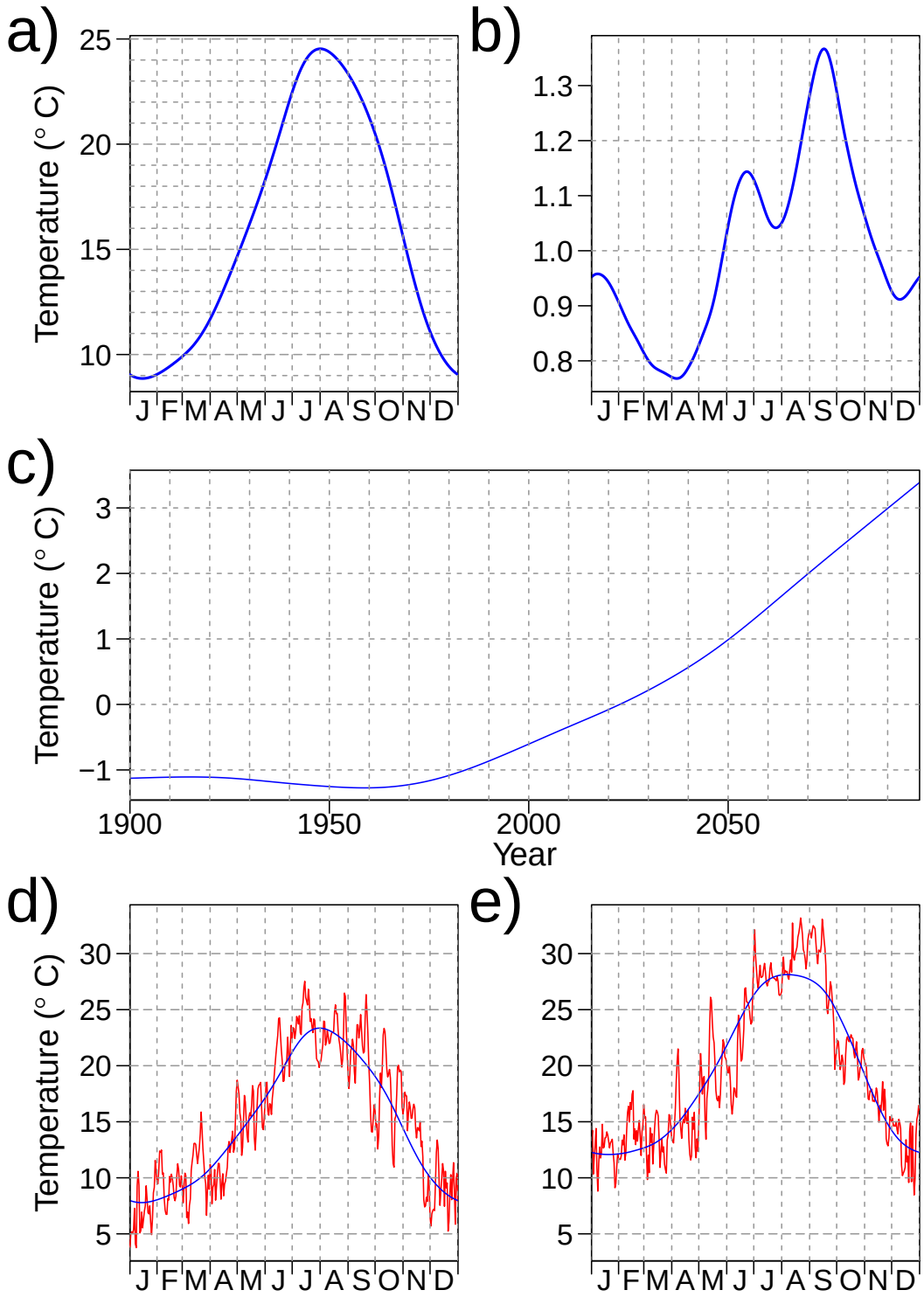


FIGURE 3.10 – Decomposition of a time series (San Francisco) by the spline model (9) trained on the 1900-2099. a) represents the reference seasonal cycle f with *df=11.5*, b) illustrates the seasonal drift h with *df=12*, and c) represents the annual trend g with *df=6*. The plots d) and e) show the estimation of the annual cycle in 1900 and 2100 respectively. Raw data is shown in red, while the fit of model (9) is in blue.

3.2 Compléments

Ces compléments se placent dans une optique d'analyse du changement climatique, par opposition à la prédiction de la normale de l'année suivante. L'objectif sera donc d'obtenir un modèle explicatif et non prédictif (comme développé précédemment). Nous présentons différents modèles, paramétriques et non-paramétriques satisfaisants à cette étude - par abus de langage, dans notre cas, cela signifiera ne pas être un sous-modèle du modèle bilinéaire. La plupart des modèles paramétriques s'inscrivent dans le cadre du modèle multiplicatif (4.1). Ceux-ci permettent, en plus d'une estimation du cycle annuel, d'apprécier la forme du changement climatique au cours de l'année.
Il est à noter que les différentes méthodes sélectionnées produisent une information similaire en terme de changement climatique, que ce soit l'amplitude ou la distribution de ce réchauffement (au cours de l'année), montrant ainsi la robustesse de nos résultats. Nous trouvons une évolution saisonnière du réchauffement pouvant doubler au cours de l'année ; c'est le cas pour Toulouse entre la période hivernal et estivale.

3.2.1 Outils de Modélisation et d'analyse

Nous rappelons dans cette section le modèle statistique d'intérêt, que nous appellerons "Modèle 0 (non-implémenté), ainsi que les sous-modèles implémentés. De plus, nous présentons des modèles de krigeage, ne partageant pas la forme multiplicative des modèles précédents. Ces derniers ont été testés sur deux types de données : les séries chronologiques de températures journalières de deux-cent-cinquante-et-une années (avec les moyennes, minimales et maximales) et, dans une moindre mesure, les données d'aires et de volumes de la banquise Arctique (trente-sept années) A.1.

3.2.1.1 Modèle 0 : modèle statistique général du "pattern scaling"

Nous exposons ici le modèle statistique général qui est inspiré de l'article de Azaïs & Ribes, 2016 [15] où la variable expliquée $T_{d,y}$ s'écrit comme la somme de deux fonctions f et $g.h$, explicitées plus bas.

$$\boxed{T_{d,y} = f(d) + g(y).h(d) + \varepsilon_{d,y}}$$

où

- $d \in [\![1, p]\!]$ représente les jours et $y \in [\![1, n]\!]$ les années.
- ε vecteur gaussien de matrice de variance covariance $\sum = \sum_g \bigotimes \sum_h$.
- f, h sont périodiques.
- f, g, h sont régulières.

f représente alors un cycle annuel de référence. $g(y).h(d)$ un terme de dérive climatique où l'on fait l'hypothèse que ce dernier se factorise en un cycle annuel h quantifiant la forme du changement climatique multiplié par g l'amplitude de ce changement.
Dès que $n \neq p$ une estimation simultanée de $\sum_g, \sum_h$ est difficile et n'a que peu d'applications pratiques ([15]), on suppose donc $\sum_g$ connue. Finalement, il nous faut estimer $f, g, h, \sum_g$.
Étant principalement intéressés par la valeur moyenne, nous ne considérerons pas l'estimation de la matrice de covariance qui sera fixée à $\sum = \sigma I_{np}$.

Remarque 3.1
Il a été montré [15] qu'il était impossible d'estimer les paramètres de ce modèle par maximum de vraisemblance lorsque qu'aucune contrainte de régularité n'était imposée sur les fonctions f, g, h.

3.2.1.2 Sous-modèles du modèle général

Dans les sous-modèles de régression suivants, nous décomposons les fonctions dépendant des jours (f et h) dans la base de Fourier et utilisons différentes hypothèses paramétriques sur la tendance en année g. Dans un premier temps, nous fixons le nombre d'harmoniques $k, l \in [\![1, 182]\!]$ utilisées pour décrire les cycles annuels. Les critères de sélection utilisés seront discutés à la section suivante. Dans la suite, nous utiliserons les notations ci-dessous :

- $d \in [\![1, 365]\!]$ représente les jours et y les années.
- $\varepsilon_{d,y} \sim N(0, \sigma)$ IID

- $k, l \in [\![1, 182]\!]$

Modèle 1 : Régression linéaire

Dans ce modèle, la variable expliquée est calculée comme la somme d'une tendance quadratique en année ($g(y)$ dans le modèle 0) et d'un cycle annuel (f) décomposé dans la base de Fourier discrète (modèle additif).

$$\boxed{T_{d,y} = (b_2.y^2 + b_1.y + b_0) + \sum_{i=1}^{k}(\gamma_i.cos(2.i.\pi.d) + \delta_i.sin(2.i.\pi.d)) + \varepsilon_{d,y}}$$

Remarque : Dans ce modèle, le réchauffement est supposé uniforme au cours de l'année. En effet, la différence entre le premier et dernier cycle est constante au cours de l'année.

Modèle 2 : Régression non-linéaire

Dans ce modèle $T_{d,y}$ s'écrit comme la somme d'un cycle annuel de référence (f) exprimé dans la base de Fourier discrète et d'un cycle annuel représentant la modulation du changement climatique (h) multiplié par une tendance quadratique en année (h).

$$\boxed{\begin{aligned} T_{d,y} = b_0 + \sum_{i=1}^{k}(\gamma_i.cos(2.i.\pi.d) + \delta_i.sin(2.i.\pi.d)) \\ +(c_2.y^2 + c_1.y).(\sum_{i=1}^{l}(\alpha_i.cos(2.i.\pi.d) + \beta_i.sin(2.i.\pi.d))) + \varepsilon_{d,y} \end{aligned}}$$

Modèle 3 : Spline en année puis régression linéaire

On estime g à partir des moyennes annuelles de la variable considérée, par une spline cubique de lissage puis, on se place dans le modèle linéaire suivant :

$$\boxed{\begin{aligned} T_{d,y} = b_0 + \sum_{i=1}^{k}(\gamma_i.cos(2.i.\pi.d)+\delta_i.sin(2.i.\pi.d)) \\ + g(y).(\sum_{i=1}^{l}(\alpha_i.cos(2.i.\pi.d) + \beta_i.sin(2.i.\pi.d))) + \varepsilon_{d,y} \end{aligned}}$$

Autres Modèles

Modèle 4 : Krigeage

Cette approche prend ses sources dans les travaux de Krige [86], ces derniers seront alors repris par Materon [95] pour développer les géostatistiques modernes. Ces techniques se démarquent par une conception bien particulière de la notion de fonction inconnue : la fonction y est modélisée comme une réalisation d'un processus aléatoire spatial. Dans notre cas, ce dernier sera supposé gaussien. Nous chercherons alors la loi de la variable d'intérêt conditionnée par les observations. Nous modélisons la variable explicative T en utilisant un Krigeage simple, c'est-à-dire, comme un processus gaussien stationnaire du second ordre de moyenne connue (i.e. de moyenne constante et de covariance invariante par translation) sur le bord latéral d'un cylindre i.e. :

$$\begin{array}{rccc} T: & (\Omega,\mathcal{F},\mathbb{P})\times\mathbb{S}^1\times[0,1] & \longrightarrow & \mathbb{R} \\ & (\omega,d,y) & \longmapsto & T(\omega,d,y) \end{array} \quad \text{où } (\Omega,\mathcal{F},\mathbb{P}) \text{ un espace probabilisé}$$

tel que $\forall n\in\mathbb{N}, \forall((d_1,y_1),\ldots,(d_n,y_n))\in(\mathbb{S}^1\times[0,1])^n$
la fonction $\omega\longmapsto(T(\omega,d_1,y_1),\ldots,T(\omega,d_n,y_n))$ est un vecteur gaussien.

Ce cadre est plus général que les modèles de régressions précédents. En particulier, il n'y a aucune hypothèse de factorisation sur la fonction de régression. En revanche, la covariance est supposée (en plus de la modélisation ci-dessus) pouvoir être écrite comme un produit de covariance sur chaque dimension considérée. D'autre choix auraient été envisageables (e.g. somme de covariance, covariance 2D) mais cette hypothèse reste naturelle vis-à-vis du problème étudié. L'hypothèse de périodicité est directement issue du choix de définition de la covariance sur le cylindre comme une restriction d'une covariance définie sur $\mathbb{R}^3$.

On effectue alors un Krigeage avec matrice de covariance de la forme :

$$C((y_1,d_1),(y_2,d_2)) = cov(T_{y_1,d_1^1,d_1^2}, T_{y_2,d_2^1,d_2^2}) = f_{\theta_1}(|y_1-y_2|).f_{\theta_2}(|d_1^1-d_2^1|).f_{\theta_3}(|d_1^2-d_2^2|)$$

où

- pour $i\in[\![1,3]\!]$, f_{θ_i} est la fonction de covariance paramétrée par θ_i.

- $d_1 = (d_1^1, d_1^2), \quad d_2 = (d_2^1, d_2^2) \in \mathcal{S}^1$ représente les jours placés sur le cercle unité. Plus précisément, à tout $d \in [\![1, 365]\!]$, on associe ses coordonnées dans $\mathbb{R}^2$ $(cos(\frac{2\pi.d}{365}), sin(\frac{2\pi.d}{365}))$.

- y_1, y_2 représentent les années.

Nous supposerons de surcroît que les observations de T sont entachées d'un bruit (indépendant de T) de variance σ^2 connue.

Le théorème de conditionnement Gaussien nous permet alors d'obtenir, en particulier, la loi de T en un nouveau point (y, d) conditionnellement
aux observations $\mathcal{T} = \{(T_1, (y_1, d_1)), \ldots, (T_n, (y_n, d_n))\}$:

$$\mathbb{E}\left[T_{(y,d)}|\mathcal{T}\right] \sim N\Big(\mu + k(y,d)^\top (K + \sigma In)^{-1}(\mathbf{T} - \mu), C((y,d),(y,d)) - k(y,d)^\top (K + \sigma In)^{-1} k(y,d)\Big)$$

où

- $k(y,d) = (C((y,d),(y_1,d_1)), \ldots, C((y,d),(y_n,d_n)))^\top$ et $K = (C((y_i,d_i),(y_j,d_j)))_{i \in [\![1,n]\!]}$

- μ représente l'espérance de T et $\mathcal{T} = \left(T_{(y,d)}\right)_{i \in [\![1,n]\!]}$.

Le choix du noyau est crucial, d'autant plus lorsqu'aucune tendance déterministe n'est spécifiée. De plus, pour être admissible, les noyaux doivent être définis positifs. Nous nous placerons donc dans une famille de noyaux répondants à ces critères puis estimerons les paramètres par validation croisée.

Remarque 3.2

Ces techniques de Krigeage rejoignent les techniques issues de la théorie des RKHS (pour l'estimation de la moyenne), le noyau sera alors donné par la matrice de covariance. Plus de développement sur ce point de vue peut être trouvé dans [134] chapitre 2 et 6 ou encore dans [7].

En plus du krigeage, d'autres méthodes à noyaux ont été implémentées. Leurs estimations étant inférieures, nous les avons omises dans ces compléments.

3.2.1.3 Méthodes de résolution

Régressions

Les méthodes de régression linéaire admettent une forme fermée bien connue qu'on résout par méthode directe. Ceci n'est pas toujours le cas pour des modèles non-linéaires et, c'est pourquoi nous avons préféré une approche utilisant des méthodes itératives, du type Gauss Newton, pour effectuer l'ajustement de ces modèles.

Package *Non-Linear Squares (NLS)* sous R

Nous avons utilisé le package NLS sous R qui permet d'estimer les paramètres d'une régression (possiblement non-linéaire) par la méthode des moindres carrés en utilisant l'algorithme de Gauss-Newton. Cet algorithme est décrit comme suit :

Étant donné le couple d'observation (y_i, x_i) $i \in [\![1 : n]\!]$ où les $x_i \in \mathbb{R}^m$ sont les vecteurs des variables explicatives et y la variable expliquée, on suppose que l'on peut écrire y comme une fonction de x dépendant des paramètres θ. i. e :

$\boxed{y_i = f(x_i, \theta) + \varepsilon_i \ \forall i \in [\![1 : n]\!]}$ où $\varepsilon \sim N(0, \sigma)$ IId

On cherche à estimer le paramètre $\theta \in \mathbb{R}^p (p < n)$ par la méthode des moindres carrés i.e

On minimise $\boxed{S(\theta) = \sum_{i=1}^{n} (r_i(\theta))^2 = \sum_{i=1}^{n} (y_i - f(x_i, \theta))^2}$

Nous cherchons donc les $\theta \in \mathbb{R}^p$ tels que (CN) :

$$\forall j \in [\![1 : p]\!] \ \frac{\partial S}{\partial \theta_j} = 2.(\sum_{i=1}^{n} r_i(\theta).\frac{\partial r_i(\theta)}{\partial \theta_j}) = 0$$

On a alors recours à l'algorithme itératif de Gauss-Newton pour résoudre cette dernière équation. Notons $\theta^s, s \in \mathbb{N}$ les itérés (θ^0 étant le point initial) et posons

$$\begin{array}{rccc} r : & \mathbb{R}^p & \longrightarrow & \mathbb{R}^n \\ & \theta & \longmapsto & (r_i(\theta))_{1<i<n} \end{array}$$

ainsi que $J_r(\theta) = (\frac{\partial r_i(\theta)}{\partial \theta_j})_{\substack{1 \le i \le n \\ 1 \le j \le p}} = -(\frac{\partial f(x_i, \theta)}{\partial \theta_j})_{\substack{1 \le i \le n \\ 1 \le j \le p}} \in \mathbb{M}_{n,p}(\mathbb{R})$ la matrice jacobienne de r.

Ensuite, nous calculons à chaque itération $s \in \mathbb{N}$:

1. $\delta\theta \in \mathbb{R}^p$ vérifiant les équations normales : $(J_r^T(\theta^s).J_r(\theta^s)).\delta\theta = -J_r(\theta^s).r(\theta^s)$ et

2. $\theta^{s+1} = \theta^s + \delta\theta$.

Remarque : L'algorithme de Gauss-Newton peut être vu comme une approximation linéaire de la fonction r. En effet, d'après le théorème de Taylor :
$0 = r(\theta_{min}) \approx r(\theta^s) + J_r(\theta^s).(\theta_{min} - \theta^s)$ où θ_{min} point d'annulation des résidus.
Il vient alors $-r(\theta^s) \approx J_r(\theta^s).(\theta_{min} - \theta^s)$.

Inconvénients

Un inconvénient majeur de cet algorithme est de ne converger que dans un voisinage de la solution et qu'il est très sensible au conditionnement de $(J_r^T(\theta^s).J_r(\theta^s))$. En effet, la méthode échouera dès que le rang de J_r est inférieur à p. Par exemple, il n'y a pas de convergence avec le modèle 2 (comme l'illustre la figure 3.11) sur 251 années de moyennes journalières de température (en rouge), l'estimation faite par le modèle étant en bleu.

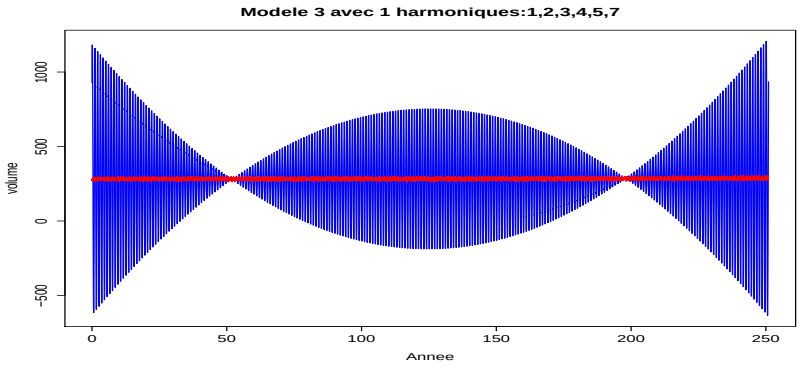


FIGURE 3.11 – Exemple de non convergence pour l'algorithme de Gauss-Newton pour le modèle 2.

Package Non-Linear Marquardt (NLMRT) sous R

Pour s'approcher suffisamment de la solution des moindres carrés, nous utilisons l'algorithme de Levenberg-Marquardt qui est plus stable que l'algorithme de Gauss-Newton et converge même s'il a été initialisé très loin du minimum.

Le point essentiel de la méthode de Levenberg-Marquardt est "d'amortir" l'équation de l'étape 1 de l'algorithme de Gauss-Newton. En effet, la matrice $(J_r^T(\theta^s).J_r(\theta^s))$ peut être mal conditionnée.

L'algorithme s'écrit comme suit :

1. Pour un $\lambda \in \mathbb{R}$

   Nous calculons $\delta\theta \in \mathbb{R}^l$ vérifiant les équations :

   $(J_r^T(\theta^s).J_r(\theta^s) + \lambda.diag(J_r^T(\theta^s).J_r(\theta^s))).\delta\theta = -J_r(\theta^s).\delta\theta$

2. $\theta^{s+1} = \theta^s + \delta\theta$

Krigeage

Le Krigeage étant une méthode d'estimation non-paramétrique, elle a permis d'évaluer la qualité des hypothèses faites dans les modèles de régression.

Package DiceKriging

Les fonctions de covariances K utilisées sont consignées dans le tableau suivant. Elles s'expriment en fonction de $h = |x - y|$ (hypothèse de covariance stationnaire) et des paramètres σ, θ, p (σ^2 est habituellement appelé le paramètre de variance et θ la longueur de corrélation). *Reg K* est mis pour "Régularité de K" et *Reg. Moy. Quad.* pour "Régularité en Moyenne Quadratique".

	Expression	Reg. K	Reg. Moy. Quad.
Gaussienne	$\sigma^2.\exp(-\frac{(\frac{h}{\theta})^2}{2})$	C^∞	C^∞
$Matern^{\frac{5}{2}}$	$\sigma^2.(1+\sqrt{5}.h/\theta + (1/3).5.(h/\theta)^2).\exp(-\sqrt{5}.h/\theta)$	C^4	C^2
$Matern^{\frac{3}{2}}$	$\sigma^2.(1+\sqrt{3}.h/\theta).\exp(-\sqrt{3}h/\theta)$	C^2	C^1
Exponentielle	$\sigma^2.\exp(-h/\theta)$	C^0	C^0
Exp. généralisée	$\sigma^2.\exp(-(h/\theta)^p)$		

Remarque 3.3

Soit Z un processus gaussien tel que pour $k \in \mathbb{N}$, sa fonction moyenne $m \in C^k$ et sa fonction de covariance $K \in C^{2k}$, alors Z est k fois dérivable en moyenne quadratique.

3.2.1.4 Critères de sélection et erreur test

Sélection *forward*

L'algorithme *forward* à été implémenté pour effectuer une sélection des harmoniques (principalement du modèle 1) avec AIC, BIC ou PRESS pour critères de sélection.
L'algorithme utilisé est le suivant :

1. Initialisation : $Liste = \emptyset$ (liste des harmoniques sélectionnées)
2. Pour K allant de 1 à 182
   (a) On rajoute à $Liste$ l'élément de $\{[\![1 : 182]\!] \setminus Liste\}$ qui améliore le plus le critère utilisé ou le détériore le moins.
   (b) On conserve la $Liste$ ayant le meilleur résultat vis-à-vis du critère.

Cependant, cette méthode à été abandonnée dans un deuxième temps à cause du temps de calcul devenant rédhibitoire notamment dans les modèles où était requise la sélection des harmoniques du cycle annuel de référence f et du delta cycle (cycle annuel du changement climatique) h.
Par exemple, sélectionner parmi les trente premières harmoniques du modèle 1 demande d'estimer le modèle 465 fois.

Sélection croissante

La sélection précédente étant trop lourde à mettre en œuvre, nous avons alors sélectionné parmi les sous-modèles, des modèles de régression contenant les $k_1 \in [\![1 : 182]\!]$ premières harmoniques pour le cycle de référence et les $k_2 \in [\![1 : 182]\!]$ premières harmoniques pour le cycle de changement climatique - celui qui minimise notre critère de sélection (AIC, BIC ou PRESS).

Algorithme utilisé :

- Pour k_1 allant de 1 à 182
  - Pour k_2 allant de 1 à 182

1. On calcule le critère utilisé sur le modèle estimé avec les k_1 premières harmoniques pour le premier signal et les k_2 premières harmoniques pour le deuxième signal.
2. On conserve le couple (k_1, k_2) ayant le meilleur résultat.

PRESS pour les données de température

On applique le critère de validation croisée PRESS (ou encore 10-fold de la section 2.3) à 10 intervalles consécutifs I_k de longueur 25 ans, formant une partition des 250 années de données de températures s'étendant de 1851 à 2100.

- $PRESS = \sum_{i=1}^{10}(\sum_{(y,d)\in I_i}(T_{(y,d)} - \hat{T}_{(y,d),-I_i})^2)$

où

- $\hat{T}_{(y,d),-I}$ est l'estimation de T en utilisant toutes les données sauf les 25 années contenues dans I.

Éléments servant à comparer les différents modèles

Voici les différents éléments calculés afin d'établir une comparaison entre les différents modèles testés sur un même échantillon E de taille $Card(E)$.
Dans la suite :

- $\bar{T}$ désigne la moyenne des $T_{d,y}$
- $\hat{T}_{d,y}$ est l'estimation faite par le modèle au jour d à l'année y

1. L'erreur quadratique moyenne :
   $MSE = \sum_{(y,d)\in E}((\hat{T}_{d,y} - T_{d,y})^2)/Card(E)$
2. Le coefficient de détermination R^2 :

$$R^2 = 1 - \frac{\sum_{(y,d)\in E}(\hat{T}_{d,y} - T_{d,y})^2}{\sum_{(y,d)\in E}(T_{d,y} - \bar{T})^2}$$

3. NbParam : Nombre de paramètres du modèle.

4. Harmo1 : Nombre de paramètres dans les premières harmoniques (cycle de référence).

5. Harmo2 : Nombre de paramètres dans les deuxièmes harmoniques (delta cycle).

3.2.2 Données et Résultats

3.2.2.1 Températures journalières

Les données de températures au pas de temps quotidien s'étalent de 1850 à 2100. Elles sont issues de la concaténation de deux simulations du modèle climatique CNRM-CM5. La première est dite historique et s'étend sur les années 1850-2005 et la deuxième est issue d'un scénario d'évolution de la concentration de gaz à effet de serre durant le $21^{ième}$ siècle RCP8.5 (c.f section 1.6). En plus de notre jeu de données principal (températures moyennes journalières), nous nous sommes aussi intéressés aux données de températures maximales et minimales.

Point de grille autour de Toulouse

Nous avons, dans un premier temps, testé nos modèles statistiques sur le point de grille le plus proche de Toulouse.

Température moyenne

Pour les moyennes journalières, le sous-modèle du modèle 0 le plus performant est le modèle 3 comme le montre le tableau 3.1 :

Régression	MSE	R^2	AIC	BIC	PRESS	NbParam	Harmo1	Harmo2
Modèle 1 BIC	10,721	0.7385	477341,8	477454,9	10,9746	11	4	0
Modèle 1 AIC	10,706	0.7388	477277,3	477692,0	10,9916	43	20	0
Modèle 1 PRESS	10,724	0.7383	477363,9	477458,1	10,9703	9	3	0
Modèle 2 BIC	10,647	0.7403	476720,8	476890,5	10,9039	17	3	4
Modèle 2 AIC	10,615	0.7410	476606,7	477530,3	10,9441	97	19	28
Modèle 2 PRESS	10,647	0.7403	476720,8	476890,5	10,9039	17	3	4
Modèle 3 BIC	10,579	0.7419	476133,6	476293,8	10,6152	16	3	4
Modèle 3 AIC	10,553	0.7425	476038,4	477801,9	10,6538	80	19	20
Modèle 3 PRESS	10,579	0.7419	476133,6	476293,8	10,6152	16	3	4

TABLE 3.1 – Tableau récapitulatif des modèles de régression 1, 2 et 3 sur les températures moyennes.

Cependant, pour tous les modèles, la sélection des harmoniques par le critère AIC est trop permissive et présente des résultats contenant trop d'oscillations, ce qui n'est pas le cas du critère BIC.

En effet, on remarque qu'on se trouve dans une situation de sur-apprentissage pour les modèles

sélectionnés par AIC. Ils offrent une meilleure erreur d'apprentissage mais de moins bons résultats concernant l'erreur de test, comme le montre la colonne contenant la valeur du PRESS. Les sélections de variables explicatives à l'aides des critères BIC ou PRESS ont donc été préférées.

On peut voir sur la figure 3.12 suivante que minimiser le critère BIC avec un choix d'harmoniques croissantes n'est pas ambigu (ce qui n'est pas toujours le cas avec le critère AIC).

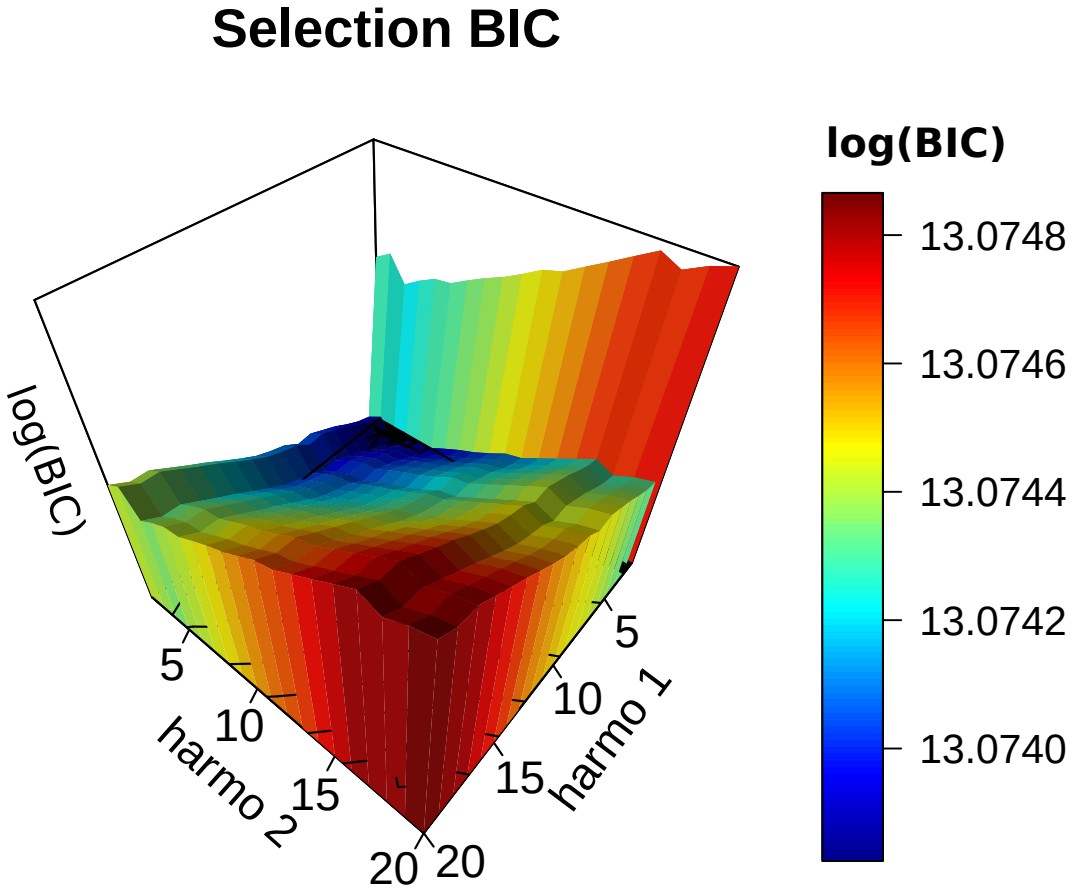


FIGURE 3.12 – Le logarithme du critère BIC en fonction du nombre d'harmoniques du cycle annuel de références (Harmo1) et du delta cycle (Harmo2). Le nombre d'harmoniques représentées varient entre 2 et 20. Le minimum du critère est atteint en (3,4).

La figure 3.13 montre la prédiction du modèle sur toute la série chronologique. On peut observer la tendance en année estimée par le modèle 3 ainsi qu'une augmentation progressive de l'amplitude du cycle annuel.

Outre le réchauffement moyen entre le premier et le dernier cycle annuel, on remarque un changement de forme. La modulation du réchauffement climatique est contenue dans le

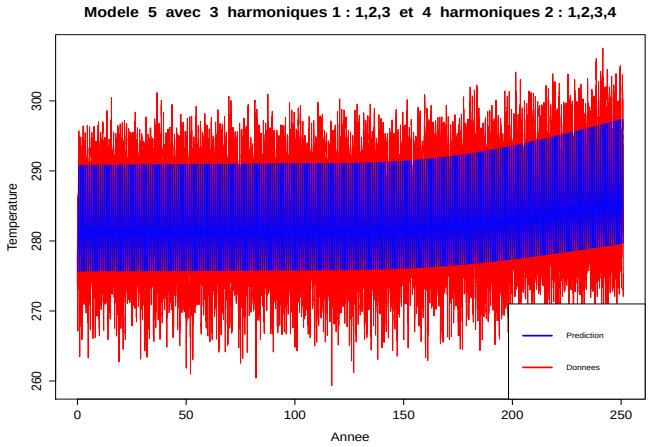


FIGURE 3.13 – modèle 3 sélectionné par critère BIC. Estimation en bleu, données en rouge.

delta cycle qui est la différence entre le premier et dernier cycle annuel. Dans la figure 3.14 le delta cycle admet un minimum global au début du mois d'avril ainsi qu'un maximum global au milieu du mois d'août. Cela signifie que le réchauffement climatique est moins fort en fin d'hiver ($3°C$) et plus fort en fin d'été ($7°C$). Cela se traduit sur la déformation des cycles annuels par un retard de l'hiver et de l'été.

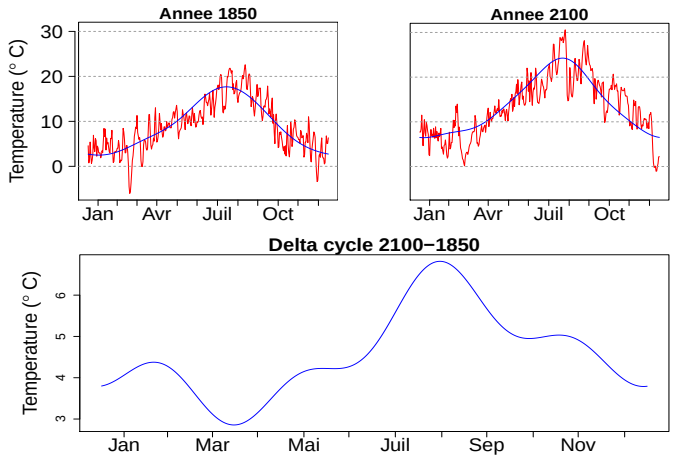


FIGURE 3.14 – modèle 3 sélectionné par critère BIC. Estimation en bleu, données en rouge.

Le diagramme quantile-quantile ci-après montre le défaut de normalité, notamment les résidus qui présentent une distribution légèrement dissymétrique dans ses quantiles les plus extrêmes.

Le graphe des résidus montre des stries verticales dues au fait que le cycle annuel "s'attarde"

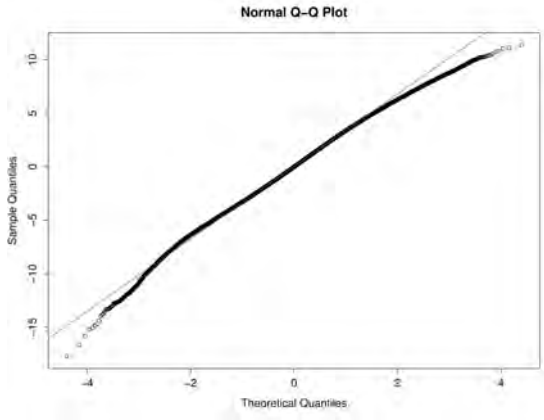


FIGURE 3.15 – QQplot.

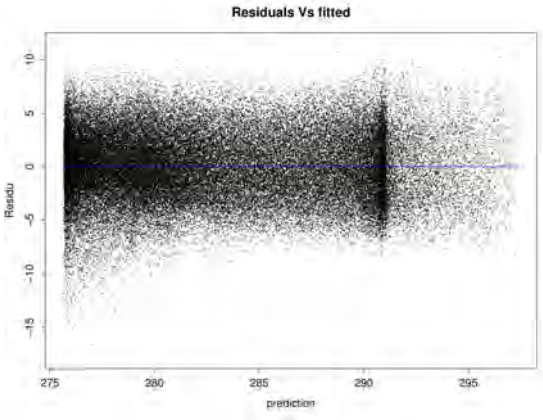


FIGURE 3.16 – Graphe des résidus.

au voisinage des minima et maxima annuels, ainsi qu'une légère augmentation de la variance pour les température froides.

La taille des données journalières étant trop importante pour implémenter les modèles de Krigeage, nous avons dû réduire notre échantillon. En effet, sans plus d'hypothèse sur la matrice de covariance (e.g nullité de la covariance passé une certaine distance), cette dernière devient difficile à inverser. Pour ce faire, nous avons tout d'abord remplacé les données quotidiennes de températures par des moyennes sur 5 jours ; ceci nous a permis d'avoir un jeu de données plus facile à exploiter. Le cycle annuel étant régulier, cette opération induit peu de perte d'information. Puis, nous avons échantillonné ce dernier tous les trois éléments afin d'obtenir notre échantillon d'apprentissage plus léger. Le reste des moyennes sur cinq jours constituant notre échantillon test. Cette opération est effectuée sur toute les données concernant les températures du climat Toulousain (minimales, maximales et moyennes).

Comme nous pouvons le voir sur le tableau 3.2, le modèle de krigeage à l'aide d'une matrice de covariance exponentielle est le modèle qui améliore le plus les critères de MSE et de R^2 sur l'échantillon test. En revanche, cette estimation ne correspond pas à une description lisse du cycle annuel. En effet, ces estimations, bien que continues, n'ont pas la régularité requise. A cet égard, les estimations à l'aide d'une matrice de covariance de Matèrn sont plus satisfaisantes. Leur erreur de test est alors très proche de celle du modèle 3 estimé sur le même jeu de données.

Régressions	MSE Test	MSE Train	R^2 Train	AIC	BIC	NbParam	Harmo1	Harmo2	
Modèle 2 AIC	6.860	7.050	0.81116	29285.8	29379.8	13	3	2	
Modèle 2 BIC	6.860	7.050	0.81116	29285.8	29379.8	13	3	2	
Modèle 3 AIC	6.792	6.960	0.81356	29213.8	29328.0	16	3	4	
Modèle 3 BIC	6.805	6.987	0.81284	29225.2	29299.1	10	3	1	
Krigeage	MSE.Test	MSE.Train	R2.Train	theta jourX	theta jourY	theta Année	p jourX	p jourY	p Année
Gaussienne	6.807	6.947	0.81390	1.49	1.39	129.00			
Matern$\frac{5}{2}$	6.803	6.939	0.81412	2.45	2.19	229.06			
Matern$\frac{3}{2}$	6.800	6.924	0.81453	2.68	2.45	298.05			
Exponentielle	6.790	6.766	0.81876	2.52	2.01	323.94			
Exponentielle généralisée	6.801	6.931	0.81432	2.84	1.98	312.85	2	2	1.856

TABLE 3.2 – Tableau récapitulatif des modèles sur les températures moyennes avec un jeu de données divisé par 15.

Résultats pour un Krigeage avec fonction de covariance Matèrn $\frac{3}{2}$

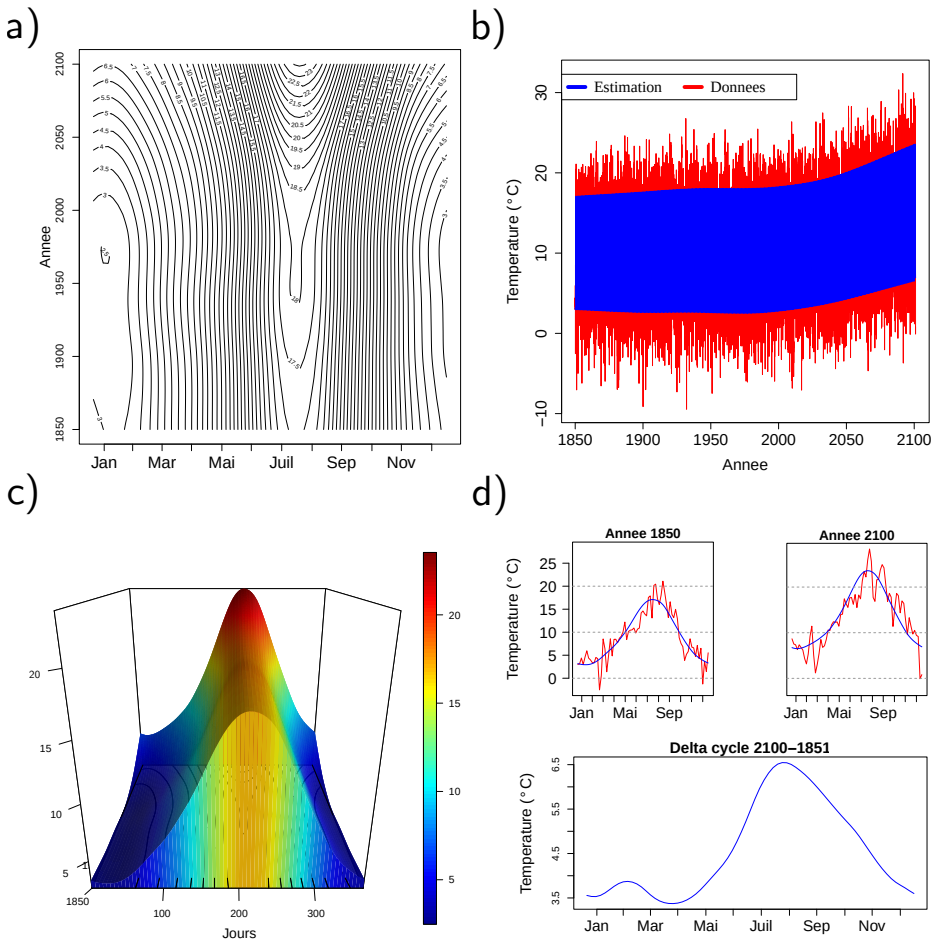


FIGURE 3.17 – Les panels a), b) et c) montrent l'intégralité de l'estimation faite par le modèle de krigeage avec matrice de covariance Matèrn 3/2. Le panel d) montre l'estimation faite pour la première et dernière année de la période considérée ainsi que leurs différences. Les panels a) et c) montrent une l'estimation 2D sous forme de contour pour le panel a) et de graphe 3D pour le panel c). Le panel b) est la représentation habituelle sous forme de série chronologique. Les données y figurent en rouge et l'estimation en bleu.

Dans la figure précédente, nous pouvons voir que le Krigeage avec une fonction de Matèrn 3/2 offre une estimation plus lisse que le modèle 3, en ce sens que la variation de la dérivée seconde des cycles estimés semble plus régulière. En terme de diagnostique sur les résidus, nous n'avons constaté aucun changement de comportement sur le graphique des résidus. Le diagramme Quantile-Quantile quant à lui montre un meilleur ajustement de la loi normale. Les données étant constituées de moyennes sur 5 jours, ce dernier résultat (amélioration

de la normalité des résidus) était attendu. Des résultats similaires peuvent être obtenus en considérant les minimales et maximales de température. Ces derniers sont reportés dans l'annexes A.2.

Les performances en terme d'erreur test du modèle 3 et 4 sont similaires. Nous noterons toutefois que la description des cycles dans la base de Fourier produit des points d'inflexion peu réalistes sur le delta cycle, mais cela n'a que peu d'impact, au premier ordre, sur la vision globale du changement saisonnier. Ce comportement peut être dû, d'une part à un choix de complexité discrète et, d'autre part, au choix d'harmoniques croissantes.

Europe avec des moyennes de températures journalières

À titre d'illustration du potentiel de ces méthodes, nous avons implémenté le modèle 3 sélectionné par critère BIC sur l'Europe entière. Ceci afin de pouvoir observer, d'une part, les décalages des dates des minima et des maxima globaux du cycle annuel et, d'autre part, les dates des minima et des maxima globaux du delta cycle. Les données utilisées pour calculer ces cartes sont des moyennes sur 5 jours, l'information du décalage est donc pertinente à cinq jours près.

Décalages de maximum et minimum

La date du maximum du delta cycle (c-à-d la date à laquelle le réchauffement saisonnier est le plus fort), reportée dans la figure 3.20, a un comportement bimodal - hormis pour quelques localisations situées en altitude. En effet, pour l'ouest et le sud de l'Europe, c'est la période estivale qui subit le plus de réchauffement. A contrario, pour le nord-est de l'Europe c'est l'hiver qui réchauffe le plus. Ce dernier point est à rapprocher de la raréfaction de la couverture neigeuse [99] en hiver. L'ensemble de ces résultats peut être retrouvés, à une échelle temporelle moins fine, dans la littérature. Par exemple, dans les rapports de l'agence européenne pour l'environnement (AEE) ou encore les études sur les données E-OBS [56] `c.f:https://www.eea.europa.eu/data-and-maps/figures/observed-temperature-change-over-europe-1976-2006`.

La date du minimum du delta cycle, figure 3.21 est plus contrastée. Néanmoins, deux grandes tendances s'en dégagent :

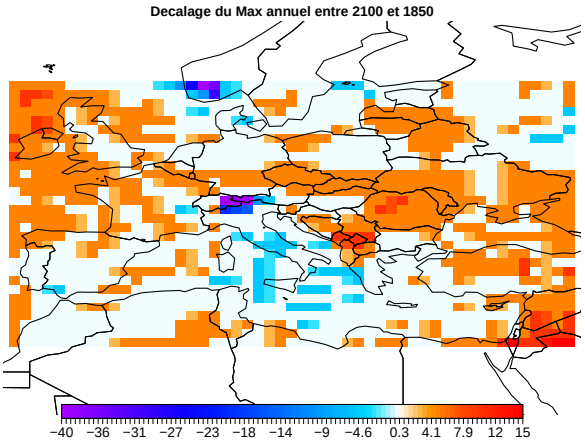


FIGURE 3.18 – Décalage de la date du maximum du cycle annuel.

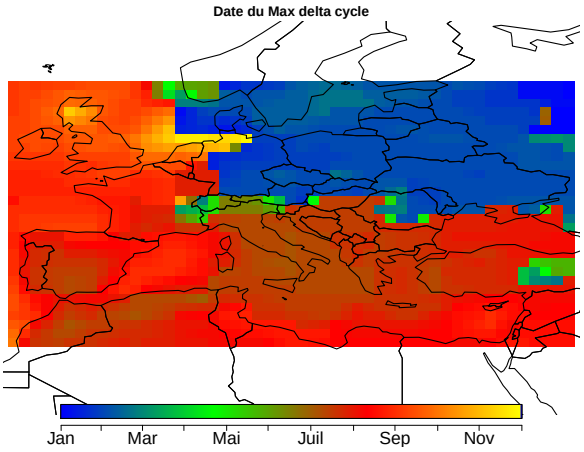


FIGURE 3.20 – Date du maximum du delta cycle.

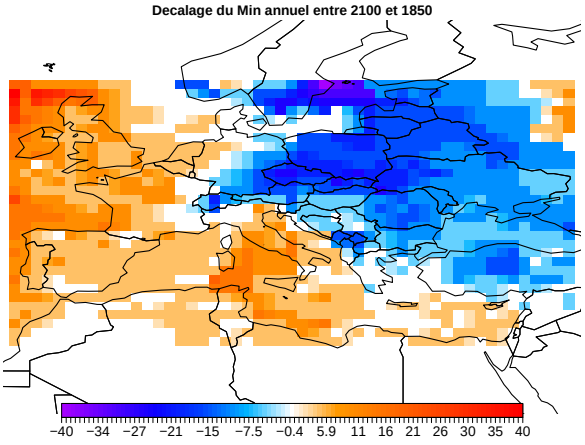


FIGURE 3.19 – Décalage de la date du minimum du cycle annuel.

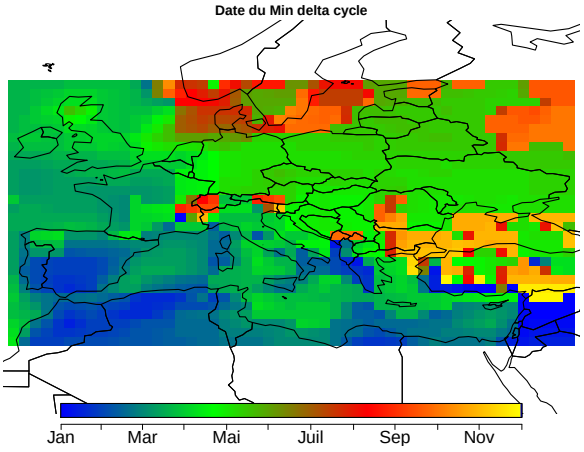


FIGURE 3.21 – Date du minimum du delta cycle.

- sur le pourtour méditerranéen, le début d'hiver (janvier) semble être la période la moins impactée, et,

- sur presque tout le reste de la zone étudiée, la partie du cycle annuel soumise à un moindre réchauffement se situe au début du printemps (avril).

En plus de cela, une troisième catégorie existe dans les hautes latitudes couvrant une partie de la zone et des plaines russes, ainsi que certaines zones à fort relief. Elles expérimentent un réchauffement mineur en été en comparaison à leur réchauffement hivernal. Ici encore nous retrouvons, avec une précision temporelle très fine, des résultats bien connues des projections climatiques régionales sur la zone Européenne [124][68].

Il est difficile de déterminer à première vue si le signal spatial du décalage du maximum

du cycle annuel est significatif (voir figure 3.18). En effet, le signal est faible et disparate (constitué de plusieurs petits motifs). Ces signaux pourraient provenir d'une structure de covariance spatiale des températures entre autres. Pour avoir un meilleur aperçu de la situation, il faudrait, par exemple, utiliser des méthodes de tests multiples, telles que les méthodes basées sur le *false discovery rate* (FDR) développé par Wilks [133, 132]. Le cas du décalage du minimum annuel (voir figure 3.19) est, en revanche, beaucoup mieux tranché : l'Europe peut être divisée en deux parties sur l'axe nord-ouest. Pour la partie à l'est, le minimum du cycle annuel tend à arriver prématurément en fin de période. Cette avance allant de 5 jours au sud-est et pouvant aller jusqu'à 35 jours dans la mer Baltique. Pour la zone située à l'ouest, l'opposé se produit aux mêmes échelles.

Conclusion

Durant ce chapitre, nous avons étudié la déformation des cycles saisonniers de l'espérance de variables climatiques, considérées au pas de temps quotidien, sous l'influence du changement climatique. Pour mesurer la plus-value de la modélisation proposée, nous avons majoritairement porté notre attention sur les températures moyennes quotidiennes.

Afin d'observer des changements dans ces phénomènes saisonniers, nous attachons une grande importance à l'étude des delta cycles (cf section 1.5). Ces derniers contiennent la modulation moyenne du changement climatique au cours de l'année. Ils permettent, entre autres, d'obtenir des diagnostics sur la différence entre les cycles de deux périodes distinctes. Dans le but d'obtenir une description parcimonieuse ainsi qu'une évolution régulière du changement, nous avons choisi d'étudier la déformation du cycle annuel au "premier ordre" ; la dérive climatique est alors modélisée comme une dilatation du delta cycle dont la magnitude est contrôlée par une fonction lisse dépendant des années. Cette décomposition multiplicative de la réponse a déjà fait ses preuves sur des décompositions de type espace-temps. Le modèle est alors mieux connu sous le nom de "pattern scaling".

C'est donc tout naturellement que nous avons été amenés à étudier les problématiques des normales tant sur le plan prédictif que sur le plan de l'analyse d'un changement moyen saisonnier au cours du $20^{\text{ème}}$ et $21^{\text{ème}}$ siècle. Cette étude est appliquée à des données modèles vues

comme une réalisation vraisemblable du climat. Pour mener à bien ce travail, les techniques de lissage issues de la théorie des RKHS (développée dans la section 2.1) étaient particulièrement appropriées et des techniques de sélection de modèles se sont avérées nécessaires pour obtenir une prédiction robuste. La complexité des modèles est alors, majoritairement ajustée par des techniques de validation croisée.

A l'aide de notre modélisation parcimonieuse de la réponse en fonction des variables temporelles (jour et année), nous trouvons que le changement saisonnier n'est pas uniforme pour toutes les localisations étudiées. En effet, les delta cycles des normales ne sont pas, pour la plupart, constants. Ce changement peut être très dépendant de la localisation spatiale considérée ; on retrouve, par exemple, un changement saisonnier plus prononcé dans les hautes latitudes et plus modéré dans les basses latitudes (section III.1.7.2).

Les travaux sur la prédiction de normales pour l'année suivante (i.e. entraînée jusqu'à l'année N pour prédire l'année N+1) vont dans le sens de la bibliographie de référence [92, 129, 12]. En effet, une meilleure prise en compte de la déformation du cycle annuel moyen, dans le but de prédire celui de l'année suivante, sont l'objet de leurs travaux. Notre méthode permet une adaptation plus flexible notamment vis-à-vis des variations séculaires et, par conséquent, offre une meilleure prédiction au cours du $21^{ème}$ siècle (section III.1.7.1). En ce qui concerne l'analyse du changement, le modèle multiplicatif semble être aussi performant que des modèles ne partageant pas cette hypothèse (section3.2). Lorsque la complexité est adaptée, on obtient une description fine et régulière du changement du cycle annuel sur la période considérée. Le modèle multiplicatif permet alors de retrouver, à une échelle temporelle plus fine (5 jours), des résultats connus à l'échelle de la saison, sur l'évolution saisonnière en Europe.

Chapitre 4

Étude de la déformation de la distribution d'une variable en climat non-stationnaire

Sommaire

Le chapitre précédent s'intéressait à l'évolution de l'espérance d'une variable climatique en fonction du temps. Comme rappelé dans le chapitre introductif, au-delà de la seule valeur moyenne (normales), le climat, considéré à un site et à une date donnée, se caractérise par une distribution de valeurs possibles. Un prolongement naturel de l'estimation de normales consiste à estimer l'ensemble de cette distribution, avec une contrainte de régularité sur la forme de celle-ci. Nous proposons alors une forme de régression quantile régularisée. On obtient ainsi, pour un paramètre donné, une description fine du climat en un site donné, et de son cycle annuel.

Dans les sections suivantes, nous allons mettre à profit les méthodes de régression quantile développées dans la section 2.2. Comme évoqué au cours de cette dernière, il est possible de tirer parti de la théorie des RKHS (e.g. on est bien dans le cadre du théorème du représentant 2.2) mais cela donne lieu à des programmes d'optimisation coûteux à résoudre. Nous préférons des pénalités conservant les programmes d'optimisation linéaire, même si pour cela nous perdons la structure de RKHS, ou encore utiliser la régression quantile sur des bases adaptées au problème (e.g. Fourier, spline). Dans les deux cas, la complexité sera estimée à l'aide des critères de sélection de la section 2.3. La pénalité la plus répandue pour la régression quantile est celle de la variation totale de la fonction. Celle-ci admet une généralisation en dimension supérieure, c'est donc assez naturellement que cette technique sera mise en place pour offrir une comparaison non-paramétrique dans l'article en préparation (section 4.1).

Dans la section 4.1 nous offrons une analyse comparant différentes méthodes existantes à notre modèle, appliquées à des observations de températures et de précipitations journalières. Nous trouvons que les méthodes de régression quantile offrent de meilleurs résultats sur ces jeux de données, notamment en tirant avantage d'une forme plus contrainte, inspirée d'hypothèses de "pattern scaling". Nous supposons en effet que la déformation des quantiles est non-uniforme au cours de l'année mais que la modulation de la dérive climatique est constante. D'autre part, nous exigeons que la magnitude de cette déformation soit gouvernée par l'évolution des températures moyennes annuelles.

Au prix d'une sélection de modèles effectuée pour chaque quantile, nous obtenons des changements significatifs dans l'évolution des distributions. Par exemple, nous trouvons une augmentation à Paris, allant jusqu'à 6 mm pour les précipitations extrêmes en été. La mise

à profit de ces résultats en suivi climatique, ne pourrait se faire sans l'aide de diagnostics efficaces. Nous en proposons deux : le delta cycle, qui est un diagnostic analogue à celui utilisé pour les normales mais appliqué à plusieurs quantiles, et un diagnostique utilisant le transport optimal univarié permettant d'observer les non-linéarités dans le changement de la distribution.

En complément, nous proposons d'étudier un modèle pénalisant la norme infinie de la dérivée seconde (i.e contraignant le maximum de la courbure) de chacune des fonctions composant le modèle multiplicatif sur des données de températures. Dans ce dernier modèle, nous relâchons une hypothèse faite par les modèles précédents. En effet, la magnitude du changement ne sera pas déterminée de façon uniforme sur les quantiles par le changement des températures moyennes annuelles.

4.1 Estimating daily climatological distribution in a changing climate

Alix Rigal, Jean-Marc Azaïs, Aurélien Ribes

Article en préparation

Résumé

Nous proposons une nouvelle méthode pour estimer la distribution de variables climatiques journalières dans un climat non-stationnaire. Le cadre statistique choisi repose sur une hypothèse de variation lisse de chaque quantile de la réponse, en fonction des variables temporelles, ainsi que sur une décomposition de la réponse inspirée par des hypothèses de "pattern scaling". L'estimation est menée à bien, en faisant usage de méthodes de lissage et de sélection de modèles, dans le cadre de la régression quantile. La complexité des modèles sera l'objet d'un examen approfondi. La méthodologie proposée sera comparée aux précédentes alternatives, telles que des modèles de régression paramétrique permettant d'observer un changement de la moyenne et de la variance de la distribution de la variable d'intérêt. Les résultats sur les données de précipitations et de températures montrent que notre technique surpasse les développements les plus récents sur le sujet sur une partie non-négligeable de la distribution, avec de surcroît une forme plus interprétable et parcimonieuse. Nous pensons que l'acquisition de telles distributions sera très utile pour la compréhension du changement climatique et trouvera de multiples applications dans le domaine le suivi climatique.

Abstract

We propose a new method to estimate daily climatological distribution in a non-stationary climate. Our statistical framework relies on the assumption that the response of each quantile to climate change is smooth over time and on a decomposition of the response inspired by the pattern scaling assumption. Estimation is carried out in the quantile regression framework using smoothing techniques, with a careful examination of the selection of smoothing parameters. The new method is compared to previously proposed alternatives such as parametric

location-scale fit and more recent methods over quantile regression splines, on corrected observations. Results on precipitation and temperature data show that our technique outperforms the most recent development with a more constrained and interpretable form. We argue that "climate change corrected" distributions might be very useful for climate monitoring.

IV.1.1 Introduction

It is commonly asked to characterize a given meteorological situation in light of its occurrence probability [123] and more recently to attribute how much climate change has contributed to its existence. Retrospective climate monitoring also typically involves such a comparison. Although this is especially the case on extreme events of a climate variable, the remaining parts of the distribution are also of interest even in this context (e.g. depending one's extreme event definition) [24]. Thus, obtaining knowledge on the whole distribution leads to a better understanding of climate evolution. Furthermore, stochastic weather generators are usually in need of a good estimate of the conditional daily density [131] in addition to the temporal dependencies of the variable (usually modeled by Markovian approaches).

We choose to focus on univariate daily climate variables which, taken at a given date, can be fully characterized by their probability distribution of possible values. It is particularly the concern of climate science to offer such an estimation of it. Moreover, in climate change studies, we are willing to obtain in what manner the distribution is significantly changing. In the most common estimation techniques, parametric forms are imposed to belong to a given family of distribution. One of the main drawbacks of such studies is that the imposed parametric form might not be adequate to represent all the features distribution. Furthermore, the family to consider might be dependent of the considered spatial localization. In other words, quality of the conclusion can be very dependent on the hypothesized shape of the distribution. Also, gaining more insights on localized changes of the distribution (i.e changes in small parts of the distribution), are hard to capture in those frameworks.

We will focus our attention on two climate variables, temperature and precipitation. To derive information on the daily temperature and precipitation, we will be using quantile regression techniques in much the same fashion as [55], but with a higher expectation in the

description of the seasonal drift in quantiles and on observation (i.e. using one unique time serie to derive significant results). The splines tensor product model (model 2 in this study), developed by [55] has the capacity to explain any regular signal. Yet, it is well known that when a simpler model holds, it yields better performances (and in our case it was deemed that the results of model 2 does not justify the use of a less parsimonious model).

In this paper, we assess the accuracy of some of the previously proposed techniques for the estimation of climatological distribution. We outline their limitations when applied to the observational records. We then introduce a new approach which overcomes these issues. Within this approach, the drift related to climate change on the seasonal component is estimated, leading to daily estimates.The manuscript is organized as follows: after presenting the data set and the methods used to estimate non-stationnary distributions in Section IV.1.2, we introduce our new method in Section IV.1.2. The estimations and scoring of the various techniques considered are assessed and discussed in Section III.1.4. This is followed by a discussion along with some concluding remarks in the last section.

IV.1.2 Data & existing methods

IV.1.2.1 Data

In this study, our main goal being methodological, we shall work on a unique location: in situ observations from Paris-Montsouris spanning from 1959 to 2017, which have been processed in order to account for ruptures due to changes in measurement conditions. Two variables are considered: daily mean temperature and precipitations. Concerning precipitations, values lower than 0.1 mm are set to 0.
Additionally, in Appendix IV.1.6.A, we use series of daily temperatures, simulated by a climate model, namely CNRM-CM5, from the Coupled Model Intercomparison Project Phase 5 (CMIP5) as a realistic realization of observations. This is in order to provide limitation of the SIC (see section III.1.2) selection, for estimating distribution during the 20^{th} and 21^{st} century, which is a period where observations are not available.

It must be noted that for daily calculations, all the 29^{th} February were removed to facilitate

processing.

IV.1.2.2 Existing methods considered in this study

We review different types of regression methods proposed by various authors to estimate climatological distribution. They are included as reference models for comparison to the one we propose. This list of techniques is not meant to be exhaustive but instead representative of proposed frameworks in the literature. Additionally, throughout this study, the considered quantile is denoted by $\tau \in]0, 1[$. It should be noted that when several quantiles are estimated, quantile crossing may occur. This is fixed using usual rearrangement techniques [27] posterior to estimation.

- **IV.1.2.2.1 Parametric/Basic models (model 0)**

  Numerous authors have proposed to estimate the daily mean temperature or daily precipitation distribution as a parametric one [66]. These works offer a subtle estimation of the parametric dimension of the distribution or the choice of the covariates. Nevertheless, in this work, we are mainly interested in the differences between a model of the location shift form - i.e. a model which can be parameterized by the mean and variance of each observation - with non-parametric ones.
  Thus, we propose two parametric models: a Gaussian model for temperatures and a gamma model on daily wet-day precipitations. Both are modeled within the generalized linear model (GLM) [135, 137] and vector generalized linear model (VGLM) framework [142, 141]. Estimation was carried out using an iterated re-weighted least squares (IRLS) procedure in order to maximize the likelihood (ML).

  Daily temperature

  We assume that daily temperatures follow a Gaussian distribution $T_{d,y} \sim \mathcal{N}(\mu_{d,y}, \sigma^2_{d,y})$ parameterized by its mean and variance for each day d and year y. The mean $\mu_{d,y}$ is estimated following the corks of [15, 109]. The variance is modeled as a gamma

distribution. Our statistical model assumes that the following decomposition holds:

$$\begin{aligned} \mu_{d,y} =& f(d) + g(y)h(d) \qquad d \in [\![1, 365]\!], y \in [\![1, n]\!], \\ H(\sigma^2_{d,y}) =& \alpha + \sum_i a_i f_i(d) + \sum_j b_j g_j(y) + \sum_{ij} c_{i,j} h_i(d).s_j(y) \end{aligned} \tag{4.1}$$

where for the mean component:

- $f(.), g(.), h(.)$ are smooth functions ($f(d), g(y), h(d)$ being their trace on integer values),
- $f(.)$, $h(.)$ are periodic functions with period 365.

In addition, we impose the constraints $\sum_{y=1}^{n} g(y) = 0$ and $\sum_{d=1}^{365} h(y) = 1$, in order to ensure the model's identifiability (i.e. to avoid any possible confusion between the terms f, g and h).

For the components relative to variance:

- $f_i(.)$, $h_j(.)$ denote periodic spline basis functions. It is not required that the two bases share the same cardinality.
- $g_i(.)$, $s_j(.)$ denote natural spline basis functions. It is not required that the two bases share the same cardinality.
- $H(.) = \log(.)$ is the link functions associated to variance.

The estimation is carried out sequentially. First we estimate $\mu_{d,y}$ using the smoothing splines techniques developed in [109], then we calculate the standard deviation estimate $\widehat{\sigma}_{d,y}$ by fitting a generalized additive model over the squared residuals $r^2_{y,d} = (T_{y,d} - \hat{T}_{y,d})^2$, taking the conditional distribution to be gamma and log link function. This is a natural choice as chi-squared distribution are particular cases of gamma distribution.

In other words, the estimated τ^{th} quantile of temperature $\widehat{T}_{d,y,\tau}$, of day d and year y can be written:

$$\widehat{T}_{d,y,\tau} = \widehat{f}(d) + \widehat{g}(y)\widehat{h}(d) + Q_{\mathcal{N}}(0, \widehat{\sigma}^2_{d,y}; \tau), \qquad d \in [\![1, 365]\!], y \in [\![1, n]\!], \tag{4.2}$$

where $Q_{\mathcal{N}}(0, \sigma^2; \tau)$ is the τ^{th} quantile of the normal distribution of variance σ^2.

Daily precipitation data

We suppose that the distribution of the daily precipitation $Pr_{d,y}$ is a mixture of an atom at zero with weight p (representing dry days) and a continuous Gamma distribution with weight $1-p$ (representing intensity of wet days). The associated parameterization of the wet day density is given by:
$f_{\mu,\gamma}(x) = \frac{\gamma}{\mu\Gamma(\gamma)}\left(\frac{\gamma x}{\mu}\right)^{\gamma-1}e^{-\frac{\gamma x}{\mu}}$, where x is the rain intensity and Γ the usual gamma function.

The daily precipitations were modeled in the GLM/VGLM framework within a two-step procedure in order to consider the discrete and continuous parts of the distribution separately.

We first estimated the dry days probability $p_{d,y}$ using logistic regression against the predictand $D_{d,y}$, which can take only two values: $D_{d,y} = 1$ if day d at year y had less than 0.1 mm of precipitation, 0 otherwise. Equivalently, this model falls in the class of GLM models with binomial family function and logistic link function [22].

A widely used density to model precipitation intensity is the two-parameter-Gamma-distribution [66]. This is why in a second step, we selected the two-parameter-Gamma-distribution to approximate the remaining density by calculating its mean $\mu_{d,y}$ and shape $\gamma_{d,y}$. Estimation of the Gamma density was carried out within the VGAM framework which enabled a simultaneous estimation of $\mu_{d,y}$ and $\gamma_{d,y}$. Following the work of [66] to account for small events, the rainfall intensity data was shifted by 0.1 mm towards zero, prior to estimation (thus 0.1 mm were added posterior to estimation). For all three components of the model, the regressors are tensor products of periodic regression splines (in order to account for seasonality), and natural regression splines.

$$\begin{aligned}
H_1(p_{d,y}) &= \alpha^{(1)} + \sum_i a_i^{(1)} f_i(d) + \sum_j b_j^{(1)} g_j(y) + \sum_{i,j} c_{i,j}^{(1)} h_i(d).s_j(y),\\
H_2(\mu_{d,y}) &= \alpha^{(2)} + \sum_i a_i^{(2)} f_i(d) + \sum_j b_j^{(2)} g_j(y) + \sum_{i,j} c_{i,j}^{(2)} h_i(d).s_j(y),\\
H_3(\gamma_{d,y}) &= \alpha^{(3)} + \sum_i a_i^{(3)} f_i(d) + \sum_j b_j^{(3)} g_j(y) + \sum_{i,j} c_{i,j}^{(3)} h_i(d).s_j(y),\\
&d \in [\![1, 365]\!], y \in [\![1, n]\!],
\end{aligned} \tag{4.3}$$

where:

– link functions are $H_1(p) = \log(\frac{p}{1-p})$, and $H_2(.) = H_3(.) = \log(.)$.

– $h_i(.)$, $f_i(.)$ and $s_j(.)$, $g_j(.)$ denote periodic and natural spline basis functions respectively.

Once all parameters have been estimated, we can derive an estimate of mixed quantile function $Pr_{d,y}(\tau)$ of precipitation at a given day and year:

$$\widehat{Pr}_{d,y}(\tau) = I(\tau \geq \widehat{p}_{d,y}).((Q_\Gamma(\widehat{\mu}_{d,y}, \widehat{\gamma}_{d,y}; (\tau - p_{d,y})/(1 - p_{d,y}))) + 0.1)$$

where $Q_\Gamma(\mu, \gamma; \tau)$ is the τ^{th} quantile of the gamma distribution of mean μ and shape γ.

- **IV.1.2.2.2 Penalized triogram (model 1)**

This method is similar to the well-known *thin plate spline* but in the quantile regression framework and with splines of order one. The latter can be found as the minimum of regression problems penalized by total variation of the gradient. This allows the problem to be written in the linear programming framework making it more affordable than splines with quadratic penalties such as [20]. Its non-parametric form and the fact that complexity can be prescribed in a continuous manner gives us indication on the general form of the change and can be used as a benchmark. Nonetheless, we should keep in mind that it suffers from modeling problems such as the fact that no periodicity assumption is made on the seasonal component and that the smoothness required for each quantile is only continuous. Lastly, it is computationally unfeasible, for a large number of quantiles and localisations, to estimate the smoothing parameters. In other words, this model would not meet the operational constraints imposed by climate services.

We choose to follow the minimization program from [78]:

$$\widehat{T}_{d,y,\tau} = \underset{\mathcal{G}}{\operatorname{argmin}} \sum_{y=1}^{n} \sum_{d=1}^{365} \rho_\tau(T_{d,y} - s(d,y)) + \lambda V\Big(\nabla s\Big), \qquad d \in [\![1, 365]\!], y \in [\![1, n]\!], \quad (4.4)$$

where V is the total variation of function $s : \mathbb{R}^2 \mapsto \mathbb{R}$, $\mathcal{G}$ is the space of piece-wise linear function over a given triangulation.

Recall that if s is smooth enough (i.e s has absolutely continuous gradient [74]),

$V\left(\nabla s\right) = \int ||\nabla^2 s(x)||dx$, where $||\nabla^2 s||$ denotes the Hilbert–Schmidt norm of the Hessian and ρ is the usual cost function associated with quantile regression:
$\rho_\tau(u) = u(\tau - I(u < 0)) = u(\tau - 1)I(u < 0) + u\tau I(u \geq 0)$

The minimum of such optimisation problem (4.4) is achieved by piece-wise linear functions on a triangulation of the observed (d, y).

- **IV.1.2.2.3 ANOVA decomposition [55] (model 2)**

  This model was developed by [55] to achieve better insight in climate simulation. Their method used 50 simulations of CESM climate model, assumed to be nearly independent, on the period $1850 - 2100$, to fit quantile regression on a tensor product of cubic regression spline basis. Also, two main effects, namely seasonal and secular, were added and described in the regression spline framework.

$$T_{d,y,\tau} = \alpha + \sum_i a_i f_i(d, \tau) + \sum_j b_j g_j(y, \tau) + \sum_{ij} c_{i,j} h_i(d, \tau).s_j(y, \tau) \tag{4.5}$$

  $d \in [\![1, 365]\!], y \in [\![1, n]\!]$

  where:

  - $f_i(.)$, $h_j(.)$ denote periodic spline basis functions.
  - $g_i(.)$, $s_j(.)$ denote natural spline basis functions.

This type of structured regression, better known as the analysis-of-variance (ANOVA) decomposition, could potentially explain any level of interaction but was tested on temperature data only.

IV.1.3 New Method

All methods described above, except ANOVA quantile regression splines (model 2), could be criticized for a certain lack of flexibility or even be assumed to have poor model definition.

Indeed, quantiles are either assumed to be of a given parametric family or moving linearly over time, with no periodicity assumption (triogram). In this section, we introduce an alternative method for computing climatological density taking into account climate change deformations of the distribution, which is somehow at least as flexible as previous methods. Indeed, it is based on spline smoothing and a covariate containing the yearly mean amount of change. Furthermore, as the imposed multiplicative structure coerces attention on the delta cycles, we believe the difference between two distinct years' distributions shall be better represented. At least, the first order deformation, according to the annual trend, will be better coerced.

This will be discussed throughout the investigation of the overall performance of our approach in subsequent sections.

IV.1.3.1 Statistical framework

The general statistical model considered here is inspired by and adapted from [15, 109]. Let $T_{y,d,\tau}$ be the τ^{th} temperature quantile of day d in year y. Our statistical model assumes that the following decomposition holds:

$$T_{d,y,\tau} = f(d,\tau) + g(y)h(d,\tau), \qquad d \in [\![1, 365]\!], y \in [\![1, n]\!], \tag{4.6}$$

where:

- $f(.,\tau), g(.), h(.,\tau)$ are smooth functions ($f(d,\tau), g(y,\tau), h(d,\tau)$ being their trace on integer values),
- $f(.,\tau), h(.,\tau)$ are, additionally, periodic functions with period 365.

Furthermore, we impose the constraints $\sum_{y=1}^{n} g(y) = 0$ and $\sum_{d=1}^{365} h(y,\tau) = 1$, in order to ensure the model's identifiability.

We propose that g is given by a smoothing spline on the yearly mean temperature, as a realistic proxy of the magnitude of the change. We subtract to g its mean over the con-

sidered year $\bar{g} = \frac{1}{n}\sum_{y=1}^{n} g(y)$ in order to satisfy the identifiability condition of the model. Furthermore, we explore two descriptions of the seasonal components $f(,\tau), h(,\tau)$:

- **(model 3 Fourier)**: $f(,\tau)$, $h(,\tau)$ are expressed in the Fourier basis. In other words, we have:
  $f(d,\tau) = a(0,\tau) + \sum_{k=1}^{n} \left(a(k,\tau).\cos(\frac{2.k\pi}{365}.d) + b(k,\tau).\sin(\frac{2.k\pi}{365}.d)\right)$
  $h(d,\tau) = 1 + \sum_{k=1}^{m} \left(a(k,\tau).\cos(\frac{2.k\pi}{365}.d) + b(k,\tau).\sin(\frac{2.k\pi}{365}.d)\right)$

- **(model 4 spline)**: $f(,\tau)$, $h(,\tau)$ are decomposed in a periodic regression spline basis containing an intercept. As for the description in Fourier basis, intercept of h is set to 1.

In addition, when multiple quantiles are considered, we enforce monotonicity using usual rearrangement techniques [27] on the ensemble of the estimated quantiles.

This statistical model can be interpreted as follows: $f(d,\tau)$ represents a stationary evolution of distributions along season, as would be observed if the climate was stationary. The effect of climate change is then described by the term $g(y)h(d,\tau)$. The key assumption is that this climate change response can be factorized into one component which describes how the shape of a seasonal cycle representing one quantile changes, $h(d,\tau)$, and another one which describes the annual variation of the magnitude of this change with time, in the long-term, $g(y)$. This type of decomposition is an adaptation of the *pattern scaling* assumption [96, 120, 46] in a slightly different setup. Under *pattern scaling*, it is assumed that the spatial distribution of climate change does not vary with time as only the amplitude of the change does. It is thus possible to decompose climate change as the product between one spatial function and one temporal function. In the present paper, the spatial component is replaced by the seasonal cycle. In both cases, the assumption can be thought of as a Taylor approximation of order one, which is valid as long as the change is small enough. This factorization assumption is obviously one of consequence but has already proven its descriptive capabilities on hourly surface air temperature observations [125]. Its primary interest comes from the induced reduction in the model's complexity: estimating two univariate functions g and h is more easily achieved than estimating a bivariate function (say $c(y,d)$). Its introduction therefore allows

us to better constrain the estimation of the climate change component. If this assumption were invalid, the predictive performance of our method would be reduced, in particular if compared to methods not relying on a similar assumption. Lastly, model 3 and 4 proved a very good capability across the entire time series considered.

An illustration of this model and the typical outputs it can produce is delayed to section III.1.4 in Figure 4.2. The R-scripts used to carry out estimation of model 3 and 4 are available online via the CNRM-GAME website (URL: `http://www.umr-cnrm.fr/spip.php?article1064`).
Next, we will discuss how the complexity is acquired within each model and the metrics used to compare them to one another.

Assessment of model complexity and bootstrap

In order to benefit from the best performances of a model, one has to tune its complexity. Higher complexity models enable to explain more subtle signals, but may result in overfiting. Less complex models on the other hand offer estimates with bigger bias but smaller variance. To acquire the complexity for each percentile τ, we used a variation of 10-fold cross-validation taking entire years out of the sample in order to prevent auto-correlation issues. We will note this statistic $10fold(\tau)$. Other selection techniques were studied such as hold out or Shwartz information criterion (SIC) and yielded similar results only less robust. For instance, regarding centiles higher than 95^{th} and lower than the $5th$ for temperatures (see Appendix B).
In order to compare the mean generalisation error of each model over quantiles an aggregated, hinge loss was used:

$$\overline{Err} = \frac{1}{N}\sum_{\tau} 10fold(\tau) = \frac{1}{N}\sum_{\tau}\sum_{d,y} \rho_\tau(T_{d,y} - \hat{T}^{-k(y)}_{d,y,\tau}),$$

where:

- ρ is the usual cost function associated with quantile regression:
  $\rho_\tau(u) = u(\tau - I(u < 0))$.
- $\hat{T}^{-k(y)}_{d,y,\tau}$ is the predicted quantile of day d and year y trained on the data set with the

fold containing y removed.

- N is number of considered quantiles.

Consequently an estimation that would minimize ρ_τ for each quantile τ would minimize $\overline{Err}$. However, this score measures the overall performance of the quantile estimation and does not place error of each quantile at the same level. Typically a model could perform better on one set of quantiles and still globally perform poorly when considering the entire distribution. This is why this score should be used as an indicator of the global performances rather than a determining criteria.

Inference over the estimated quantiles was made using a simple block bootstrap procedure. It was carried out in a similar fashion as cross-validation, i.e. taking entire years out of the considered times series so as to avoid dependency issues. Using this methodology, we could derive confidence intervals and tests on the nullity of the delta cycles for each quantiles. Further details can be found in Appendix IV.1.6.C as well as an application on daily precipitation.

IV.1.4 Results

Temperature

In figure 4.1, the performances in terms of mean prediction $\overline{Err}$ is quite clear for temperature. In the first column of figure 4.1, models can be ranked in increasing order according to their overall quantile performance (namely 0, 1, 2, 4 and 3). The parametric model 0 showed highest mean prediction error and also the biggest variability. Although the penalized triogram method (model 1) has stringent regularity assumption and requires no periodicity in the seasonal component, it clearly performed better than model 0. Furthermore, its aggregated score $\overline{Err}$ yields a much smaller variance than model 0. Also, for each model, improvement is nearly linear according to model number, until model 3 and 4, which perform similarly with a slight advantage for model 3, interquartile ranges are similar to model 1 (about .3). It is interesting to note that the improvement in performance, for this aggregated score, between the parametric model and model 2, is less or equal to the one between the latter and model 3.

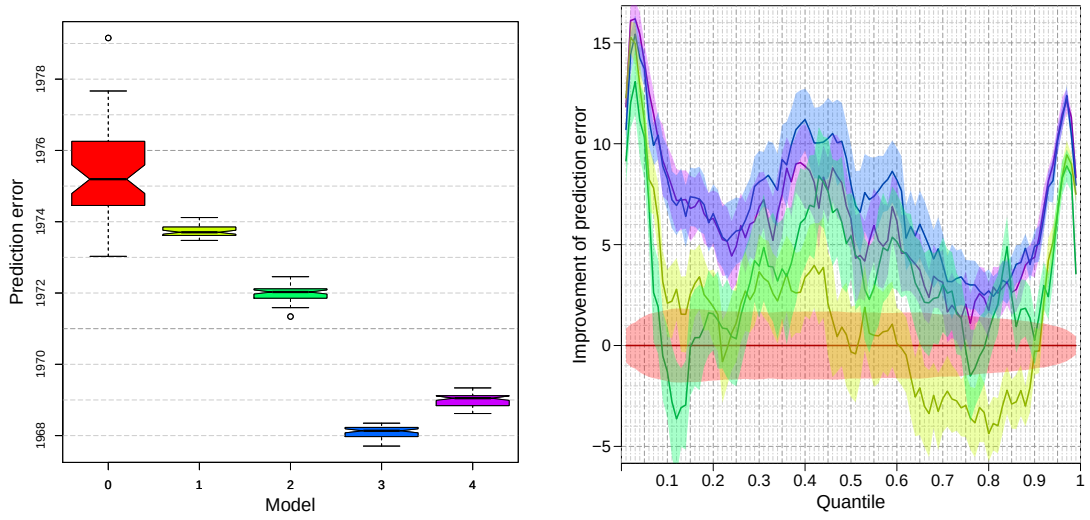


Figure 4.1 – Scoring of the five models with non-constant complexity selected by 10-fold cross-validation on mean daily temperature observations. The prediction error is also estimated using the 10-fold cross-validation statistic, by resampling over the folds. Here, 50 partitions have been chosen randomly. The left panel shows mean prediction $\overline{Err}$ over quantiles. If the notches of two boxes do not overlap, this is 'strong evidence' that the two medians differ [26]. The right panel shows mean prediction error of model 0, subtracted by mean prediction error of each model, for each quantile with their respective standard error estimates. Model 0 is represented in red, 1 is yellow, 2 is green, 3 is blue, and 4 is purple.

The right panel of figure 4.1 shows mean prediction error of model 0 subtracted by mean prediction error of each model, as a function of quantiles τ. This allows to reduce the variability of the score as prediction error has a very large variation range rendering it impossible to analyse performances according to quantile, see Appendix IV.1.6.D). Thus, each curve represents improvement of prediction error, for each model, with respect to model 0. A positive value indicating that the model yields better performances. In addition, standard deviation of the estimates have been added, in order to account for uncertainties of the scores. For the Gaussian model 0, the improvement in prediction error is, by definition, set to 0. The standard deviations are relatively constant and tend to diminish for high quantiles from .70 to .99 and low quantiles from 0.15 to 0.01. Overall quantile regression models always perform much better beneath first decile and above the 9^{th} decile, and tend to be superior on the more central part of the distribution. Model 1 and 2 yielded equal or lower performances relatively to model 0 in two regions of the distribution: between .1 to .2 and 0.65 and .9. The models based on the pattern scaling assumption, namely model 4 and 3, dominates all other alternatives. Fourier decomposition of seasonal cycles tends to be slightly more effective than

the periodic spline basis approach on the central part of the distribution, i.e. between 0.25 and 0.75. On the other hand, opposite conclusion are met below the first decile.

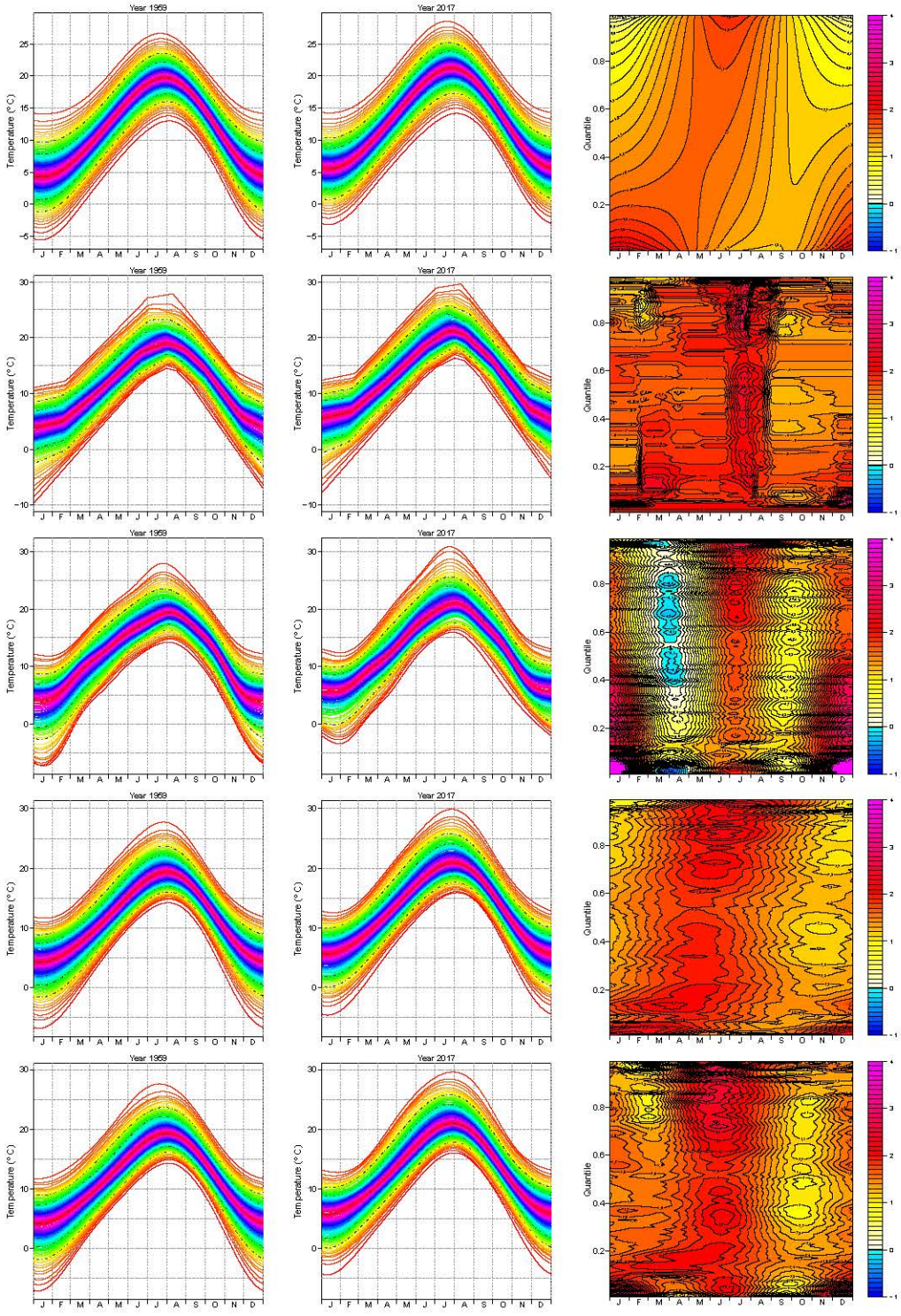


Figure 4.2 – Estimation of percentiles in Paris-Montsouris daily mean temperature for each model. From left to right: estimation for the year 1959, 2017, and delta cycles. The models are represented in increasing order from top to bottom (models 0, 1, 2, 3 and 4) with complexity acquired for each quantile. The color scale is the same for each plots. Additionally, for plots of the annual cycle, deciles are represented by dashed lines.

The first two columns of Figure 4.2 show results of estimation, for each model, on the first and last years of the period, namely 1959 and 2017. Overall no severe quantile crossing was encountered and all models but model 1 and 2 yielded very similar annual cycle. Indeed, model 1, due to its piecewise linear form, generated more quantile crossing than other models and performed poorly at the boundaries of the year. This can well be seen on the most extreme cold events of January. Model 2 showed a more complex deformation of the annual cycle which, for instance, substantially changes the curvature of the cycles in spring. Additionally, we observed consequent (and quite unrealistic) bi-modality in the estimated distributions, which can be seen, for example, by observing spacing between quantiles .85 and .86 in July - August 2017. This phenomenon also occurs on other models with discrete variation of complexity according to the quantiles but at a much smaller scale. Following the work of [114] all models but model 1 showed a strong diminution of the variance of daily mean temperatures in winter driven by the disappearing of extreme cold temperatures - a phenomenon being particularly strong in January.

The last column of figure 4.2 shows delta cycles between years 1959 and 2017 for each quantiles. Each horizontal slice thus represents the warming experienced between the last and first year, for a given quantile along the year. Qualitatively, models tend to show non-uniform warming over the daily distributions, with bigger warming on the lower quantiles in winter and on higher quantiles in summer. This pattern is very much in line with preceding studies on observed seasonal changes in Paris on temperature indices [5]. Model 0, 3 and 4 give a very smooth response and also tend to show a band that experienced more warming, extending from March to July This corresponds to a stronger change in the lower quantile at the beginning of the period which progressively spreads to the higher quantiles at the end of the period. Although model 3 is globally the winning model in terms of scoring over quantiles (figure 4.1), it is inferior to model 4 below the first decile, for instance we can see that it is having difficulties in producing more localized functions for cold extremes. Model 1 showed more vertical structures between quantiles .1 and .8, showing nearly uniform and stronger warming of the distribution in July-August, a still uniform and less pronounced warming from September to January, and a transition period on which lower quantiles experienced more change ongoing from February to June. Model 2 yielded smooth structures with a much more pronounced seasonal change compared to other models. Changes in the extreme low

distribution in winter are the biggest among models, with more than 4 degrees augmentation. On the less extreme part of the distribution, the delta cycle showed larger warming of the lower part of the distribution (nearly twice) in winter. Summer distribution yielded the opposite behavior only less pronounced, i.e more warming in the upper part of the distribution. March-April showed nearly no warming and even cooling of the distribution on the central and extreme cold part of the distribution. Finally, from September to October, the temperatures undergo moderate warming about 0.8 degree.
All quantile regression models tend to prescribe more warming than the parametric model 0 particularly on the most extreme parts of the distribution nearly one degree for cold extremes in winter, .5 for hot extremes in summer when contrasted to model 4.

The transport diagnostic

This subsection considers the question of the difference between an heteroskedastic Gaussian (model 0) with a quantile regression model on a continuous climate variable such as temperature. We propose to analyse this difference using optimal transportation on the real line. Indeed, it is long known [107] that for two distributions with no atom (or equivalently continuous cumulative distribution functions) and strictly convex cost, there is a unique map that transforms the first distribution onto the other at the lowest cost. The solution of this Monge problem can be written as a composition of quantile and cumulative distribution function (CDF). For instance, let $T_{d,1959}$, $T_{d,2017}$ be two random variables representing temperature at a given day d for years 1959 and 2017 with respective CDF, $F_{d,1959}$ and $F_{d,2017}$. Then $F^{-1}_{d,2017} \circ F_{d,1959}$ transforms $T_{d,1959}$ into $T_{d,2017}$. For the Gaussian model 0, the transport map is given by an affine transformation as $T_{d,2017}$ can be obtained by a simple translation and dilatation of $T_{d,1959}$. In order to visualize in what manner quantile regression better describes the evolution of temperature distributions, we propose to withdraw the transport map of model 0 to the transport map of model 4 for each day. In other words, for each day d we calculate $F^{-1}_{d,2017} \circ F_{d,1959} - G^{-1}_{d,2017} \circ G_{d,1959}$, where F and G CDFs are estimated by model 4 and 0 respectively. This diagnosis is shown in the right panel of figure 4.3. This is very close to the study of the residual of a linear fit (for each day) on the empirical transport map estimates.

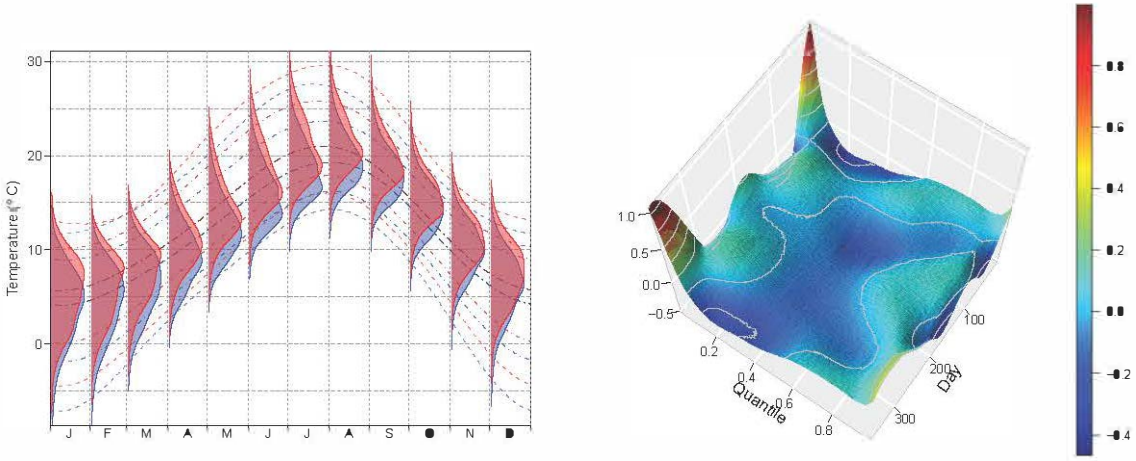


Figure 4.3 – The left panel shows estimated distributions by model 4 for the first day of each month, year 1959 in blue and year 2017 in red respectively. Additionnally shown are the evolution throughout the considered years of medians (two dashed central lines), first and last deciles and percentiles (respectively dotted and dashed lines and simply dashed lines). The right panel shows the difference between transport map of estimated distributions of model 4 and 0, for each day between 1959 and 2017.

The left panel of figure 4.3 displays the estimation of distributions made by model 4 in 1959 and 2017. In order to obtain smooth density estimates from the quantile process, adaptative kernel methods [115], which are a trade off between K nearest neighbors and kernel methods, were applied. Apart from the shift and change in variance, it is quite difficult to quantify more subtle deformations that are not explained by the heteroskedastic model 0, this is shown in the right panel of figure 4.3. The steepest non-linearity occurs on extreme cold quantiles during the end of autumn and the beginning of winter. It shows that, in addition to the warming depicted by the Gaussian model 0 in this region (already 2 degrees), model 4 adds one supplementary degree. Also, more moderated effects can be seen such as cooling of .2 degree in December and January for quantiles between .15 to .4 and cooling of about .3 degree for extreme low quantiles in April.

Precipitation

In figure 4.4, the performances in terms of mean prediction $\overline{Err}$ for precipitations yielded similar conclusions than for temperature. In the first column of figure 4.1, models can be

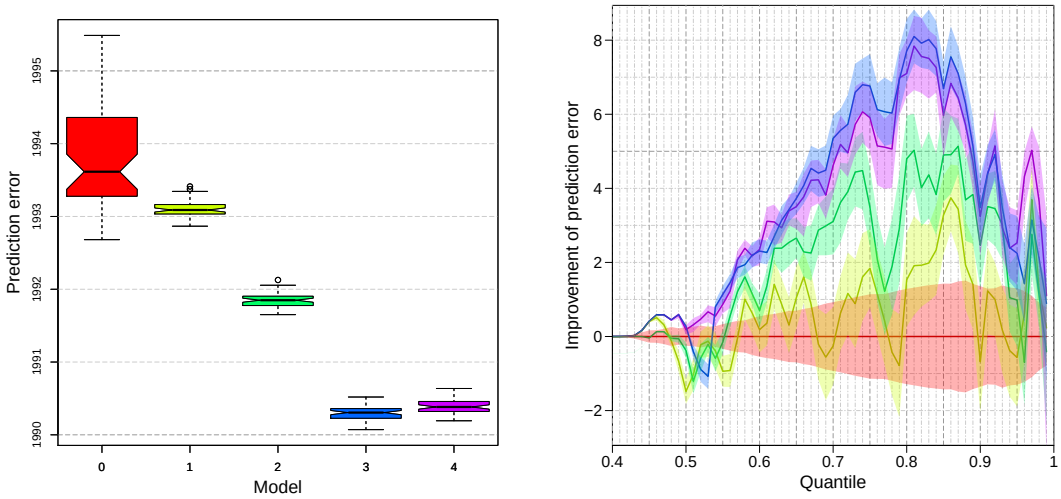


Figure 4.4 – Scoring of the five models with non-constant complexity selected by 10-fold cross-validation on precipitation observations. The prediction error $\overline{Err}$ is also calculated using re-sampled 10-fold cross-validation. The left panel shows the aggregated (over quantiles) mean prediction error. If the notches of two boxes do not overlap this is 'strong evidence' that the two medians differ [26]. The right panel shows scores minus the score of model 0 for each quantile, with standard error estimates. Model 0 is represented in red, 1 is yellow, 2 is green, 3 is blue and 4 is purple.

ranked in increasing order according to their overall quantile performance (namely 0,1,2,4 and 3). The parametric model 0 showed highest mean prediction error and, in addition, the biggest variability. Although the penalized triogram method (model 1) has stringent regularity assumption and requires no periodicity in the seasonal component, it clearly performed better than model 0. Furthermore, its aggregated score $\overline{Err}$ yields a much smaller variance than model 0. Afterwards, for each model, improvement is nearly linear according to model number until model 3 and 4 which perform similarly, with a slight advantage for model 3. Interquartile ranges are similar for all models but model 0 (about .3). It is interesting to note that the improvement in performance between the parametric model and model 2, for this aggregated score, is less or equal to the one between the latter and model 3.

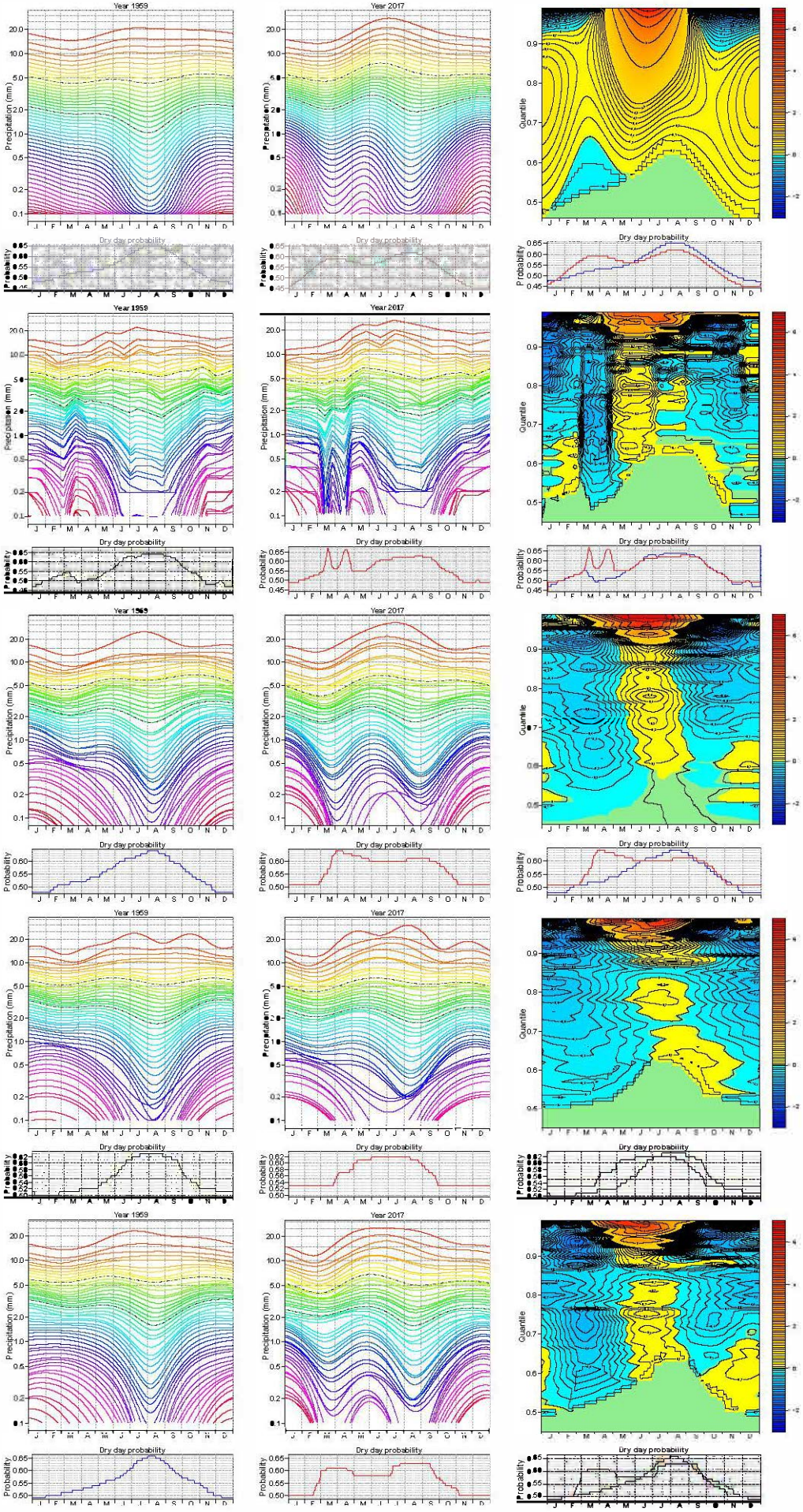
The right panel of figure 4.1 shows mean prediction error of model 0 subtracted by mean prediction error of each model, as a function of quantiles τ. This enables to reduce the variability of the score. Indeed, prediction has a very large variation range, making it impossible to analyse performances according to quantile (see Appendix IV.1.6.D). Thus, each curve represents performances of each model relative to model 0. When the improvement in prediction

error is positive for a given model, it yields better performances. In addition, standard deviation of the estimates have been added in order to account for uncertainties of the scores. For the gamma model 0, the improvement is, by definition, set to 0. The standard deviations are increasing linearly until quantile .96, then they tend to diminish for high quantiles from .87 to .99. Overall quantile regression models tend to perform better for quantiles above 6^{th} decile. On the other hand, they tend to be inferior between quantiles .45 and .6. These quantiles depict the transition between the atom (zero precipitation) and the non-zero quantiles.

Model 1 was clearly inferior to model 0 between .42 to .58 and yielded greater or equal performance relatively to model 0 on other quantiles. Model 2 yielded lower performances than model 0 on a similar region that is .45 to .56 then clearly outperformed model 0. Model 3's score deterioration on the transition from dry to wet day occurred on a smaller interval (.5,.54). It clearly outperformed all models 0, 1 and 2 on the remaining quantiles below 95. It then shared similar performances with model 1 and 2 on higher quantiles. Finally, model 4 is the only model that does not suffer a score discrepancy on the lower quantiles. Overall, it yields better scoring than all proposed alternatives below quantile 95 and apart from quantiles concerning the wet day/dry day transition. The Model 4 had the same results as model 3 with a very slight advantage for model 3 on interval .65 to 90. After quantile 95, model 4 clearly outperformed all other alternatives.

Models based on the pattern scaling assumption, namely model 4 and 3, dominated all other alternatives. Fourier decomposition of seasonal cycles tends to be slightly more effective than the periodic spline basis approach on the central part of the wet day distribution between 0.65 and 0.90, but this is at the cost of poor performances on more extreme quantiles.

Figure 4.5 – Please see description on next page.



In figure 4.5 illustrates the estimation of percentiles in Paris-Montsouris daily precipitations for each model. From left to right: estimation for the year 1959, 2017, and delta cycles. Models are represented in increasing order from top to bottom with complexity acquired for each quantile. The corresponding dry day probabilities are below each panel. Precipitation quantile estimations are log-scaled and quantile delta cycle below 45 are not represented as no change occurs. The first two columns of figure 4.5 shows estimated distributions of daily precipitation for Paris-Montsouris on the first (1959) and last (2017) year of the period. Each row represents a model. For each model, the upper panels represents the (estimated) percentiles evolution along both considered years. In addition, the lower panels show the seasonal progression of dry day probability. Qualitatively the quantiles between .7 and .95 shared the same behavior, regardless of the considered model. They show an evolution of quantile cycles, from a marked minimum in July-August for lower quantiles, to cycles with its maximum on June-July for the bigger events. In other words, summer period tend to have less small events but bigger extreme precipitations. For extreme height precipitation, the majority of models (namely models namely 0, 2, 4) showed uni-modal quantile cycles with its maximum attained in summer and minimum in winter. Model 1 though rather similar to the others, yielded two modes, most of the time. The Fourier representation of the cycles in Model 3 produced much more variation than other models, yielding more than two maxima along the year. This quite unrealistic description of extreme height precipitations given by model 3, is coherent with the poor model scoring shown in Figure 4.4 for quantiles higher than .95.
Most of the quantile crossing occurred in the lower quantiles (below .7) and, as for temperatures, model 1 experienced the most. However most of the models (all but model 3) showed the same deformation of quantile cycle from uni-modal cycle, with its minimum in summer, to bi-modal cycles, with two minima in March-April and August-September.

The variation of dry day probability are between .45 and .65. In 1959 estimation showed nearly the same pattern for all models: maximal precipitation in winter (maximal in January-February) and minimum in summer (minimum in July-August), with a monotone variation between those two poles. The evolution of dry day probability in 2017 is not that simple, but there is also a global pattern that emerges on all models but 1 and 3; dry day probability becomes bi-modal due to a strong increase in March-April. Furthermore, according to scoring

in Figure 4.4, the most realistic evolutions are given by model 4. Indeed, for 2017, it yields annual cycles with a large period from March to September, where the probability of dry day slowly fluctuates between .58 and .63. This probability is nearly stationary around .5 for the remaining months.

The last column of figure 4.5 shows the difference between quantile cycle of year 2017 and 1959, namely the delta cycle. Hence, it represents the amount of change in precipitation distribution, between 1959 and 2017, as a function of the day and probability. Each delta cycle is represented by a horizontal slice of the contour plot. The percentiles considered range from the 45^{th} to the 99^{th}, as lower quantiles are equally zero for all models. The delta cycles produced by quantile regression models show very similar signals. Globally, there is a decrease of precipitation intensity (due to climate change) on all seasons but summer, where, after a slight augmentation on quantiles below .9 (about .1 mm to .2 mm), a much bigger event can be noticed (between 2 mm and 5 mm) spanning from May to August. Although the parametric model 0 is qualitatively different from the quantile regression models - it produces delta cycles that show mainly augmentation of precipitation on most quantiles - it shares some characteristics mainly on high quantiles such as, the strong increase from May to August and decrease on the remaining year. Furthermore, if the reader is interested in the significance of such change, this question has been addressed in Appendix IV.1.6.C using bootstrap techniques on model 4. It showed that the augmentation in precipitation intensity on the strong events of summer and diminution of moderated quantiles (below 95^{th} percentiles) were non-zero (at a significance level of 0.05).

IV.1.5 Conclusion

In this article, we proposed a comparison between parametric fits and different models in the quantile regression framework, applied on a unique time serie at a daily timescale. Overall, the latter showed superior results for precipitation and temperatures. We further analyzed two variations of a more constrained model based on a pattern scaling assumption (all of them showed better quantiles estimation) which explored the description of seasonal components in two bases, namely Fourier and spline. Results show that description with the Fourier basis are slightly better on the central part of the estimated distribution but performed poorly on the more extreme quantiles compared to splines. This is due to the better localization offered by spline function (as opposed to harmonic functions). Furthermore, this model directly estimates a interesting diagnosis for upcoming climate studies i.e. the delta cycle. It contains the seasonal change for each quantile along the year (see last column of figure 4.2 and 4.5). Thus, it informs us of where the most rapid changes occur in the distribution and along the seasonal cycle. Additionally, the difference between quantile regression models and parametric location scale models can be appreciated in the light of the transport map for continuous variables such as temperature (see figure 4.3). It is important to have precise estimation of regularity for all considered models in order to yield significant results over all quantiles (see Appendix D), as regularity changes along them. Indeed, in general, the most extreme quantiles yielded more complex annual cycles than the more central ones. With a good estimation of the complexity at hand, we recovered different results on observed changes in Paris, with a higher precision on their evolution at a daily timescale.

Perspectives for this work would be a better (faster) estimation of the complexity. Admittedly, SIC criterion gave a good overall view of the complexity but was deemed less powerful than cross-validation (see Appendix IV.1.6.C). Another prospect would be to take into account extreme value theory estimations, for example, using extremal quantile regression methods [29], so as to expand distribution estimation. Additionally, more adequate and faster methods, to perform inference should be considered (we are using basic bootstrap) in order to test significant changes in distributions.

Overall, we believe that quantile regression offers a powerfull tool to seek better insight

of climate variables. This could be valued by climate services for the significant results that can be derived on the observational record. Furthermore, this framework can be transposed easily to climate model data sets. In practice, it could be of interest for the attribution of weather and climate events.

IV.1.6 Appendix

The analysis in this article has been performed using the statistical software `R`.

IV.1.6.A Performance of SIC criterion over model 3

Figure 4.6 shows estimation of quantiles with model 3, on the CNRM-CM5 model, for years 1900 and 2098. The model selection is done in order to minimize the SIC criterion, for each quantile. It can be seen that regularity for extremal quantile is quite unrealistic. Furthermore the far-fetched antimode around the median in year 2098 is imputable to the discrete choice of regularity parameters. Nonetheless, the other parts of the distribution in 2098 seems to be well-estimated.

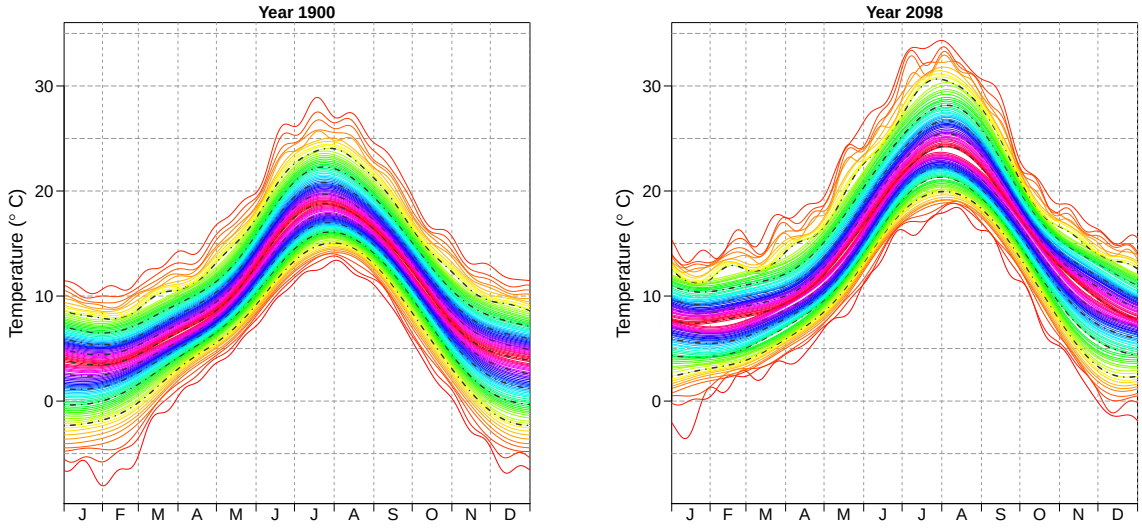


Figure 4.6 – Estimation over of the percentiles with model 3 on the entire period (1900-2098) of CNRM-CM5 with determination of the smoothness for each quantiles with SIC criterion.

IV.1.6.B Evolution of complexity for precipitation

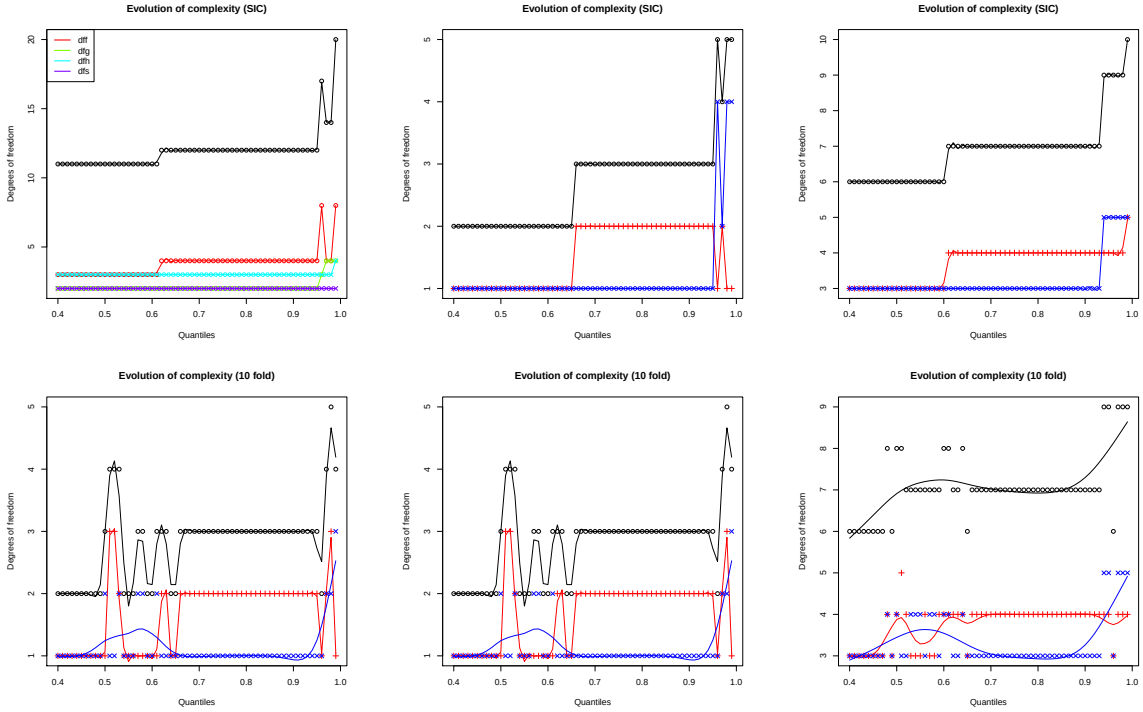


Figure 4.7 – Evolution of the selected smoothing parameters for models 2, 3 and 4, on precipitation observation, as a function of the quantiles. The top panels shows the selection using Schwarz criterion, the bottom ones using 10-fold cross-validation. Total complexity of the model is given by the black points. For models 3 and 4, the smoothing parameters required for the reference cycle and delta cycle are represented by the red and blue points, respectively. For model 2, red represents the reference cycle, green the annual trend, cyan and purple represent the number of elements in the tensor product spline basis required for the seasonal and annual components, respectively. Additionally, smoothed version of the selected complexity are provided by the curves of the corresponding colours.

IV.1.6.C Details on the inference conducted over precipitation data

After bootstrapping a 1000 times over the years, fitting the quantile regression estimate (for each bootstrap sample) for all percentile and rearranging them accordingly (for each estimates), we can obtain an estimation of the sampling variability of quantile regression. This is shown in figure 4.8. Bootstrapped confidence intervals are calculated by taking the quantiles (0.025 and 0.975) of the estimates for each year and day. Furthermore, the same bootstrap samples are used to derive the delta cycles for each quantiles. We then consider

the difference in quantile (delta cycle) to be significant when zero does not lie in its bootstrap confidence interval. Although this procedure is sub-optimal, it can clearly help us appreciate if a constant complexity over quantile offers the same results as one with model complexity that varies for each quantiles. For example, in figure 4.8, the upper panels represents the bootstrap procedure on model 4 for a constant complexity acquired by 10-fold cross validation. The lower panel represents results for the same model but with non-constant complexity over quantiles (also obtained by cross-validation). First of all, we can see that the bootstrap estimates over each deciles on the years 1959 and 2017 seems to behave similarly. Nonetheless, the bootstrapped mean delta cycle does not share the same patterns. Also the significance of the delta cycles is much higher if complexity is authorized to vary with quantile.

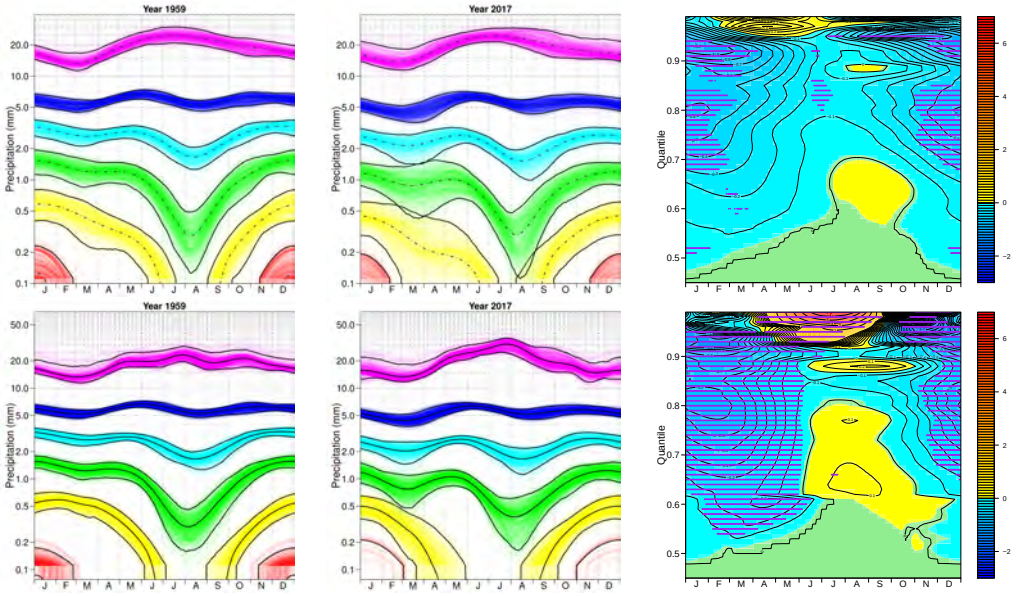


Figure 4.8 – Estimation of percentiles in Paris-Montsouris daily precipitations for model 4: from left to right estimation for the year 1959, 2017, and mean delta cycle derived from the bootstrap estimates. Purple points represents estimates that are significantly different from zero. The upper panels are derived from estimations with the same fixed smoothing parameter for all quantiles. In the bottom panels, complexity is authorized to vary with quantile. Precipitation quantile estimations are log-scaled in order to equally represent them. Quantile delta cycle below 40 are not represented as no change occurs. The color scale is the same for each plots.

IV.1.6.D Evolution of the prediction error according to quantiles

In figure 4.9, an estimation of prediction error for each model as a function of quantiles can be seen. It is hard to see which model is wining as variability of prediction error along quantiles is much bigger than variability between each model at a considered quantile τ. This is why, in figure 4.1 and 4.4, we chose to take one of the models, namely model 0, as a reference to compare performances between each models.

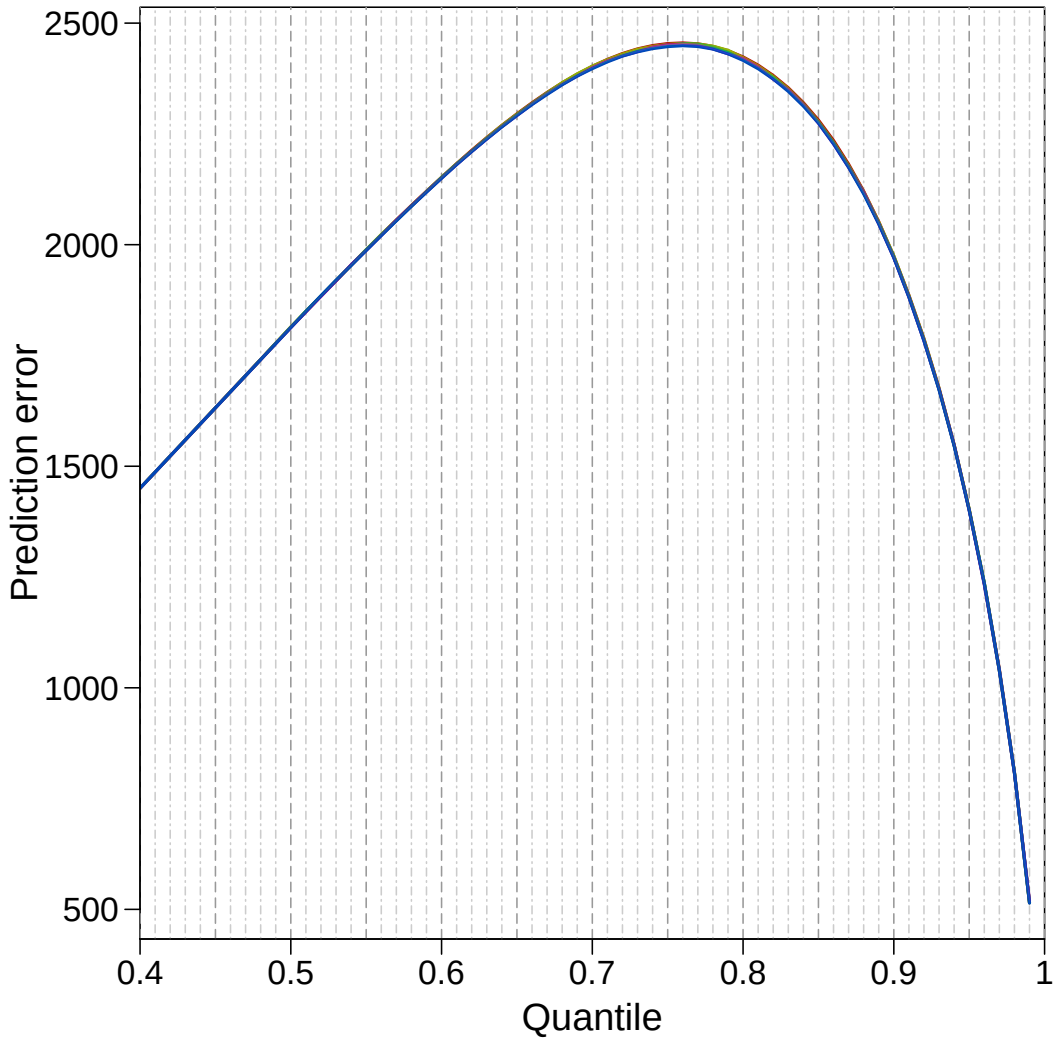


Figure 4.9 – Estimation of prediction error for each model as a function of τ, the considered quantile, on precipitation observations.

4.2 Complements

The preceding models (3 & 4) offered a better (more constrained) description of the distributions. Yet, the careful reader would have noticed that some of the hypotheses have not been tested. For example, annual magnitude of the quantile cycles deformations were uniformly distributed by function g representing the annual mean of a climate variable. However, it could be varying along the quantile considered.

Furthermore, the regularity prescription of model (3 & 4) are discrete (by opposition to continuous) which could produce artefacts such as quantile crossing or other quantitative differences in quantiles cycles. This last point is particularly true due to the strong variation of the regularity according to the quantiles. Finally, the need to estimate changing complexities depending on the considered quantiles in order to obtain better results, renders, at this point, the method unsuitable to operational constraints. To deal with two of those infelicities, we chose to generalize the preceding statistical models 3 & 4 . Let $T_{y,d,\tau}$ be the τ^{th} temperature quantile of day d in year y. Our statistical model assumes that the following decomposition holds for each quantile:

$$T_{d,y,\tau} = f(d,\tau) + g(y,\tau)h(d,\tau), \qquad d \in [\![1,365]\!], y \in [\![1,n]\!], \tag{4.7}$$

where:

- $f(.,\tau), g(.,\tau), h(.,\tau)$ are smooth functions ($f(d,\tau), g(y,\tau), h(d,\tau)$ being their trace on integer values),
- $f(.,\tau), h(.,\tau)$ are, additionally, periodic functions with period 365.

Furthermore, we impose the constraints $\sum_{y=1}^{n} g(y,\tau) = 0$ and $\sum_{d=1}^{365} h(y,\tau) = 1$, in order to ensure the model's identifiability.

We proposed, in the first two models (model 3 & 4), to take $g(y,\tau)$ as a constant function over τ, by prescribing $g(y,\tau)$ to be given by a smoothing spline on the yearly mean temperature g, as a realistic proxy of the magnitude of the change. This is no longer the case

here.

- **(model 5)**: This last model seeks to estimate every part of equation 4.7 using penalized spline techniques. We seek to solve:

$$(\widehat{f}(.,\tau),\widehat{g}(.,\tau),\widehat{h}(.,\tau)) = \underset{\mathcal{F}_p\times\mathcal{F}\times\mathcal{F}_p}{\operatorname{argmin}} \sum_{y=1}^{n}\sum_{d=1}^{365}\rho_\tau(T_{d,y} - f(d) + g(y)h(d)) + \lambda_f\|f''\|_\infty + \lambda_g\|g''\|_\infty + \lambda_h\|h''\|_\infty \tag{4.8}$$

  where:

  - $\mathcal{F}$ is the space of twice differentiable functions with continuous first derivative and $\mathcal{F}_p$ is the same space with periodicity condition (in order to model f and h).
  - $\lambda_f, \lambda_g, \lambda_h$ are smoothing parameters controlling the roughness of functions f, g, h.

  Such a minimization problem yields quadratic splines [40]. Furthermore, when estimating only one function - or even additive quantile regression models - the problem can be rewritten in the linear programming framework [79]. This virtue enables efficient computation of the estimator.

This statistical model can be interpreted as follows. $f(d,\tau)$ represents a stationary seasonal evolution of distributions, which would be observed if the climate was stationary, the effect of climate change is then described by the term $g(y,\tau)h(d,\tau)$. The key assumption is that this climate change response can be factorized into two components. One which describes how the shape of a seasonal cycle representing quantiles changes, $h(d,\tau)$, and another one which describes the variation of the magnitude of this change with time, in the long-term, $g(y,\tau)$. However, the magnitude of the change along the years is no more imposed to be uniform over quantiles.

An illustration of this model and the typical outputs it can produce is shown in figure 4.10 applied on Paris-Montsouris. It is interesting to note that the biggest mean annual change experienced occurred in lower quantile (as shown in panel c) of figure 4.10). We should note that the annual variation (panel c) shows small decadal variations on the non-central

quantiles. This could be due to internal variability or a poor estimation of the smoothing parameters. Determining the source of this behavior is beyond the scope of this section.

Next, we shall discuss how the unknown functions f, g, h are estimated within this model.

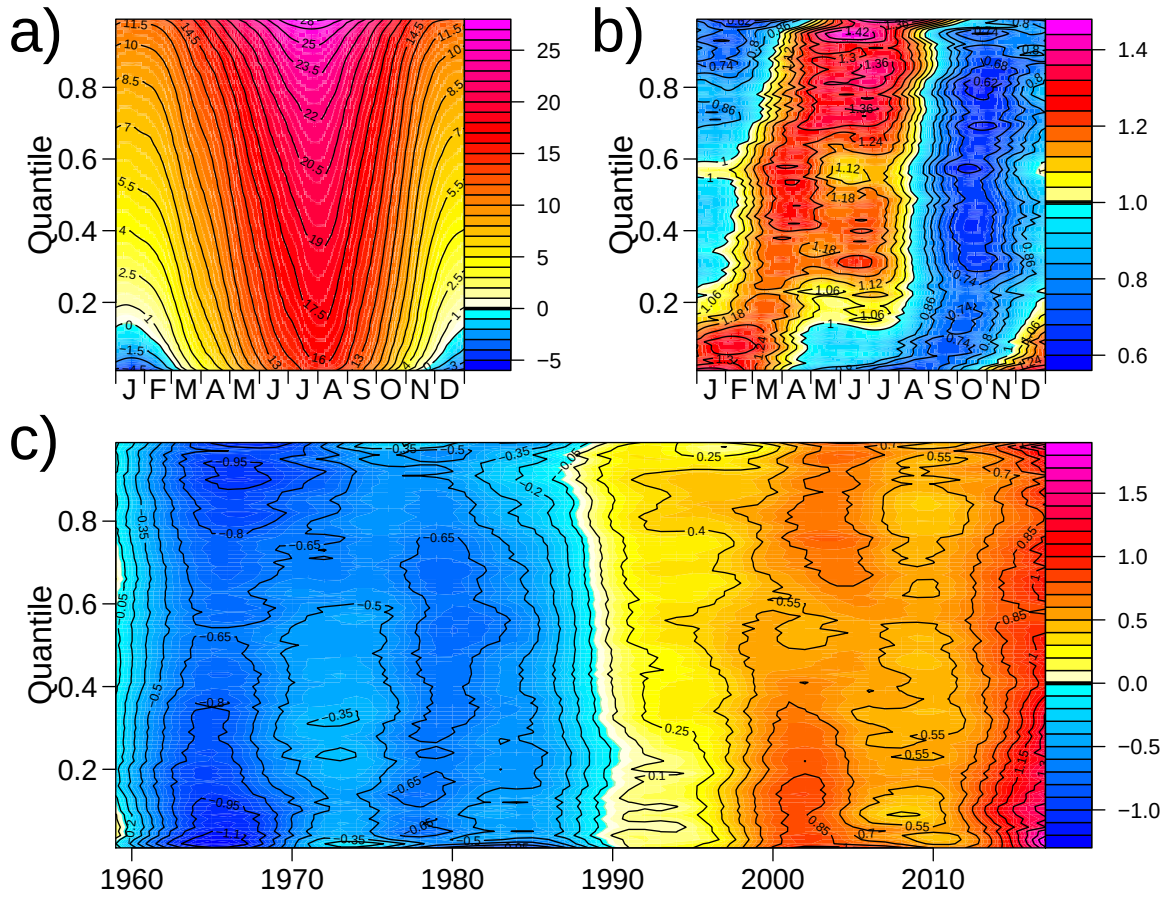


Figure 4.10 – Decomposition of a time series (Paris-Montsouris) by the penalised splines model 5 (4.8). The evolution of each quantile is represented by the evolution of the corresponding horizontal line. For each τ the complexity of $f(.,\tau)$, $h(.,\tau)$, $g(.,\tau)$ have been acquired by cross-validation (see section 2.3). a) represents the reference quantile cycles f, b) illustrates the seasonal drift h along quantiles, and c) represents the annual trend g.

4.2.0.1 Estimation Algorithm of Model 5

Furthermore, we attempted to provide a calculation which meets operational constraints and which is thus computationally not too expensive so as to be applied to multiple locations. For this reason, we implemented the sequential algorithm described below.

Algorithm: Alternating hinge loss

1) **Initialization of $g()$:**

Calculate the annual means $T_{.y}$. From the temperature $T_{.y}$ time-series, compute the smoothing spline estimate $\widehat{g}()$ of $g()$, with a given degree of freedom df_g. Note that this estimate has to be centered subsequently in order to satisfy the identifiability constrains.

2) **Cycle: until f,g,h changes less than a pre-specified threshold**

A) **estimation of $f(.,\tau)$ and $h(.,\tau)$:**

$$(\widehat{f}(.,\tau),\widehat{h}(.,\tau)) = \underset{\mathcal{F}_p\times\mathcal{F}_p}{\operatorname{argmin}} \sum_{y=1}^{n}\sum_{d=1}^{365}\rho_\tau(T_{d,y} - f(d) + \widehat{g}(y)h(d)) + \lambda_f\|f''\|_\infty + \lambda_h\|h''\|_\infty$$

B) **re-estimation of $g(.,\tau)$:**

$$\widetilde{g}(.,\tau) = \underset{\mathcal{F}}{\operatorname{argmin}} \sum_{y=1}^{n}\sum_{d=1}^{365}\rho_\tau(T_{d,y} - \widehat{f}(d,\tau) - g(y)\widehat{h}(d,\tau)) + \lambda_g\|g''\|_\infty$$

$$\widehat{g}(.,\tau) = \widetilde{g}(.,\tau) - \frac{1}{n}\sum_{y=1}^{n}\widetilde{g}(.,\tau) \text{ (centering of estimated function)}$$

As it is sequential and based on multiple evaluation of quantile regression, this algorithm yielded comparable execution time with models 3 and 4. The number of iterations needed in order to achieve convergence was between 2 and 3, depending on the considered quantile. The R-scripts used to carry out the estimation of model (4.1) are available online via the CNRM-GAME website (URL: `http://www.umr-cnrm.fr/spip.php?article1064`).

4.2.1 Temperature

The same scoring and estimation plots are displayed to compare with the methods used in the article for daily temperatures in Paris. As can be seen, model 5 offered better performances on this data set, notably on low quantiles.

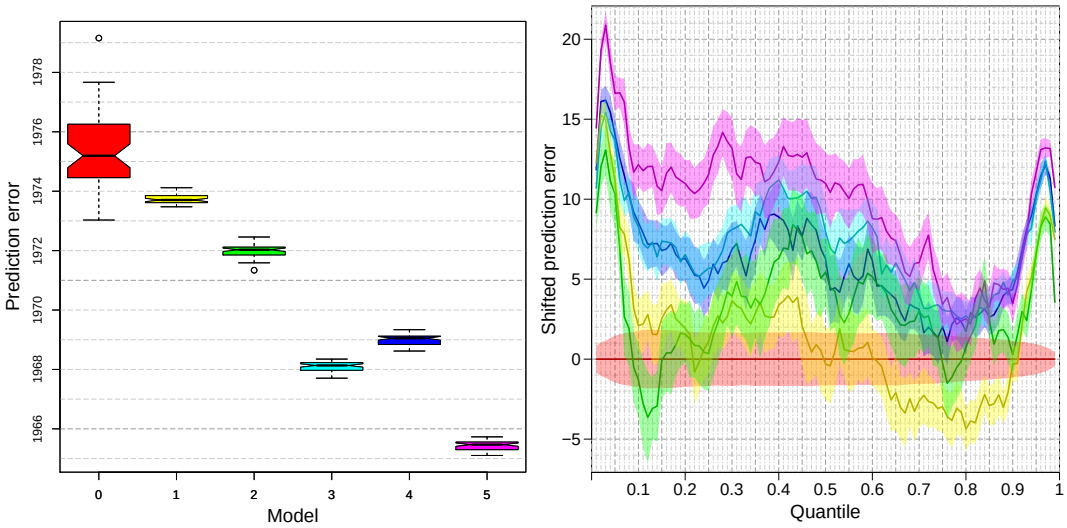


Figure 4.11 – Scoring of the five models with non-constant complexity selected by 10-fold cross-validation on mean daily temperature observations. The prediction error is also estimated using the 10-fold cross-validation statistic, by re-sampling over the folds. Here, 50 partitions have been chosen randomly. The left panel shows mean prediction Err_τ over quantiles. If the notches of two boxes do not overlap this is 'strong evidence' that the two medians differ [26]. The right panel shows mean prediction error of model 0, subtracted by mean prediction error of each model, for each quantile with their respective standard error estimates. Model 0 is represented in red, 1 is yellow, 2 is green, 3 is cyan, 4 is blue and 5 is purple.

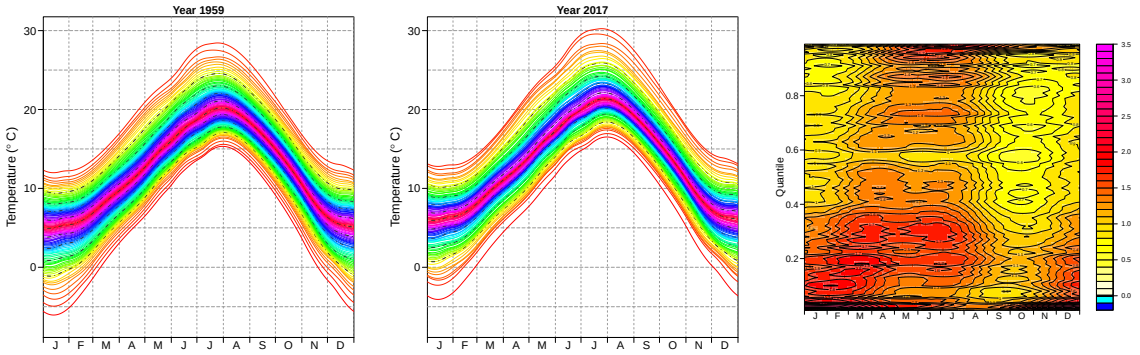


Figure 4.12 – Estimation of percentiles in Paris-Montsouris daily mean temperature for model 5. From left to right : estimation for the year 1959, 2017 and delta cycles are represented, with complexity acquired for each quantile. The color scale is the same for each plots. Additionally, for plots of the annual cycle, deciles are represented by dashed lines.

As it can be seen in figures 4.12, the evolution of quantiles depicted by model 5 are very similar to those of model 3 and 4. The delta cycles, on the other hand, are less pronounced but conserve quite the same pattern as it can be seen in the decomposition of figure 4.10. All this implies that the improvement in scoring shown in figure 4.11 is mainly due to a better estimate of the chronology of the magnitude of the change in quantile (which is governed by the function $g(y, \tau)$) on the great majority of the period.

Conclusion

Dans ce travail, nous avons proposé une modélisation permettant de décrire l'évolution des distributions au pas de temps journalier de deux variables climatiques (précipitations et températures moyennes journalières) tant sur leurs composantes saisonnières que séculaires. À notre connaissance, la modélisation proposée offre plus de finesse que les méthodes actuelles. De surcroît, nous obtenons des résultats en ne faisant usage que d'une seule série temporelle d'observations corrigées (c.f IV.1.2.1). Le problème reste néanmoins difficile car les performances de ces méthodes sont dépendantes d'une bonne estimation de la complexité des modèles (e.g Appendice IV.1.6.C). Cette dernière ne présente, à notre connaissance, pas d'à priori avantageux (c.f Appendice IV.1.6.B) tels qu'une complexité constante le long des quantiles considérés, là où même une variation régulière fait quelquefois défaut. En particulier la complexité (nombre de degrès de liberté) doit dépendre du niveau considéré. Ce qui pose des problèmes de mise en oeuvre et de temps de calcul. Cependant, des critères de sélection plus appropriés que le SIC ou la validation croisée devraient être capables d'alléger cette procédure. Ceci étant dit, nous obtenons, à l'aide de procédures bootstrap, des changements significatifs sur une majeure partie des distributions considérées, le long de leurs cycles annuels. Pour Paris, nous obtenons un fort réchauffement des extrêmes froids en hiver ainsi qu'un réchauffement moins prononcé mais plus homogène sur l'ensemble des autres saisons après le troisième quartile. D'autre part, nous obtenons deux types de signaux significatifs sur les précipitations ; une augmentation des extrêmes de précipitations dans la période estivale et une diminution globale de l'intensité des précipitations sur les autres saisons.

Conclusion et perspectives

4.1 Conclusion

Le but premier de ce travail de thèse est d'étudier la déformation des cycles saisonniers de variables climatiques, considérées au pas de temps quotidien, sous l'influence du changement climatique. Nous avons traité le sujet sous deux angles distincts. D'une part l'évolution des normales climatiques et donc de l'espérance d'une variable, par exemple la température. D'autre part, la déformation des distributions au travers de l'étude des quantiles. Pour chacune de ces alternatives nous avons proposé une méthode d'inférence originale, apportant une contribution aux méthodologies existantes.
Pour mesurer la plus value de ces modèles nous avons majoritairement porté notre attention sur deux variables offrant une bonne caractérisation du climat à une date et une localisation donnée : les températures et précipitations quotidiennes. Les données considérées sont de deux types, issues soit de modèles climatiques, soit d'observations corrigées.

Pour observer des changements dans ces phénomènes cycliques (dans notre cas saisonnier), nous attachons une grande importance à l'étude des delta cycles, cf section 1.5. Ces derniers contiennent la modulation moyenne du changement climatique au cours de l'année. Ils permettent, entre autres, d'obtenir des diagnostics sur la différence entre les cycles de deux périodes distinctes. C'est, en somme, assez naturel que la modélisation proposée prenne en compte cette composante du problème. Dans le but d'obtenir une description parcimonieuse ainsi qu'une évolution régulière du changement, nous avons choisi d'étudier la déformation du cycle annuel au "premier ordre" ; la dérive climatique est alors modélisée comme une dilatation du delta cycle dont la magnitude est contrôlée par une fonction lisse dépendant des années. Cette décomposition multiplicative de la réponse a déjà fait ses preuves sur des décompositions de type espace-temps. Le modèle est alors mieux connu sous le nom de "pattern scaling".

La première quantité étudiée est l'espérance d'une variable (chapitre 3). C'est donc tout naturellement que nous avons été amenés à étudier les problématiques des normales tant sur

le plan prédictif que sur le plan de l'analyse d'un changement moyen saisonnier au cours du $20^{ème}$ et $21^{ème}$ siècle. Cette étude est appliquée à des données modèles vues comme une réalisation réaliste du climat. Pour mener à bien ce travail, les techniques de lissage issues de la théorie des RKHS (développé dans la section 2.1) étaient particulièrement appropriées et des techniques de sélection de modèle se sont avérées nécessaires pour obtenir une prédiction robuste.

Dans un second temps, nous avons entrepris l'étude de la quasi-totalité des distributions (toujours considérées au pas de temps journalier) par le biais de l'estimation de ses quantiles. Nous avons proposé une modélisation permettant d'acquérir l'évolution des distributions au pas de temps journalier de deux variables climatiques (précipitations et températures moyennes journalières) tant sur leurs composantes saisonnières que séculaires. L'estimation est effectuée à l'aide de la théorie de la régression quantile (section 2.2). La complexité des modèles est, là aussi, un problème central. Celle-ci est ajustée par des techniques de validation croisée. Dans les deux cas, les diagnostics permettent d'apprécier le changement saisonnier basé sur le delta cycle, aident à une évaluation efficace du changement à l'échelle journalière. Dans le cas de distributions climatiques, le transport optimal 1D permet de mettre en avant le signal non-capturé par des méthodes utilisant des lois paramétriques usuelles. Cela permet de diagnostiquer les changements les plus complexes à appréhender dans les distributions.

A l'aide de notre modélisation parcimonieuse de la réponse en fonction des variables temporelles (jour et année), nous trouvons que le changement saisonnier n'est pas uniforme pour toutes les localisations étudiées. Ceci tant sur le plan de l'évolution saisonnière de l'espérance et des quantiles, que sur celui de la partie de la distribution impactée. En effet, les delta cycles des normales et quantiles ne sont pas, pour la plupart, constants. De plus, l'impact du changement sur les distributions d'un jour donné n'est pas égal le long de ses quantiles. C'est, par exemple, le cas pour les températures hivernales de Paris, où la majorité du changement est portée par la disparition des extrêmes froids. Nous obtenons des résultats significatifs de ce changement au pas de temps journalier sur des données observées et issues de modèles climatiques. Ce changement peut être très dépendant de la localisation spatiale considérée ; on retrouve, par exemple, un changement saisonnier plus prononcé dans les hautes latitudes et plus modéré dans les basses latitudes.

Les travaux sur la prédiction de normales pour l'année suivante (i.e. entraînée jusqu'à l'année N pour prédire l'année N+1) vont dans le sens de la bibliographie de référence [92, 129, 12]. En effet, une meilleure prise en compte de la déformation du cycle annuel moyen dans le but de prédire celui de l'année suivante sont l'objet de leurs travaux. Notre méthode permet une adaptation plus flexible notamment vis-à-vis des variations séculaires et, par conséquent, offre une meilleure prédiction au cours du $21^{ème}$ siècle. En ce qui concerne l'analyse du changement, le modèle multiplicatif semble être aussi performant que des modèles ne partageant pas cette hypothèse. Lorsque la complexité est adaptée, on obtient alors une description fine du changement du cycle annuel sur la période considérée.

Peu de travaux ont été menés à l'aide de la régression quantile pour approfondir l'étude des distributions climatiques à l'échelle journalière. On peut tout de même citer les travaux de [55], ces derniers utilisent 50 simulations de modèles pour obtenir des changements significatifs. Nous nous sommes donc comparés à cette méthodologie en ne faisant usage que d'une seule série d'observations et trouvons notre approche plus parcimonieuse. Nous retrouvons, de fait, différents résultats issus de précédentes études, par exemple, sur l'évolution saisonnière des extrêmes, de la variance (en hiver) et pouvons observer l'évolution régulière de ces phénomènes à une échelle plus fine que les études précédentes.

Le problème reste néanmoins difficile car les performances de ces méthodes sont dépendantes d'une bonne estimation de la complexité des modèles. Cette dernière ne présente, à notre connaissance, pas d'à priori avantageux (c.f Appendice IV.1.6.C) tels qu'une complexité constante le long des quantiles considérés, là où même une variation régulière fait quelquefois défaut. Cependant, des critères de sélection plus appropriés que le SIC ou la validation croisée devraient être capables d'alléger cette procédure. Ceci étant dit, nous obtenons, à l'aide de procédures "bootstrap", des changements significatifs sur une majeure partie des distributions considérées, le long de leurs cycles annuels. Une sélection efficace (offrant une bonne estimation de la complexité et étant algorithmiquement moins contraignante) de la régularité des modèles fait défaut à ce travail ; cela ne permet pas l'exploitation des méthodes de régression quantile sur un grand nombre de localisations sans prévoir d'attribuer une part conséquente de la charge de calcul à la validation croisée. Néanmoins, l'objet des précédents travaux étant méthodologique, il nous était impératif d'obtenir des estimations de l'erreur de généralisa-

tion. Ceci explique pourquoi cet aspect, bien qu'important au niveau applicatif, n'ait pas été étudié dans sa totalité. Ceci est le principal verrou technique pour observer les changements saisonniers à l'aide de ces modèles à l'échelle du globe, par exemple, au travers d'analogues et classifications climatiques ou encore de procédés d'analyse de données ACP. Un autre point qui n'a été abordé que sous l'égide de la performance modèle serait l'obtention de tests efficaces sur les hypothèses du modèle multiplicatif à travers le globe, par exemple, contre l'alternative d'un modèle additif (réchauffement uniforme le long de la variable saisonnière) ou encore contre un modèle moins contraint comme des splines de lissage sur le cylindre (e.g. splines 2D sans décomposition multiplicative). Cela demanderait, par exemple, une meilleure modélisation de la vraisemblance (et donc de la dépendance temporelle) notamment pour faire des tests du type ratio de vraisemblance pénalisée. Enfin, par manque de temps, nous nous sommes restreints à l'étude d'une réponse univariée.

Si cette thèse a offert quelques pistes originales que ce soit du point de vue de l'analyse du changement saisonnier ou de celui d'une prédiction des normales plus performantes, nombre de prolongements, améliorations et applications possibles de ces travaux restent à aborder et plusieurs questions demeurent ouvertes.

4.2 Perspectives

4.2.1 Espérance d'une variable climatique

Une grande partie de ce travail étant focalisée sur l'étude à fine échelle temporelle des variables, elle n'a été que peu exploitée au niveau spatial, malgré quelques résultats centrés sur l'Europe. Ceci est en parti dû à la difficulté d'estimation des paramètres de lissage des modèles. Peu de recherches, au cours de cette thèse, ont été menées dans cette direction. Il existe, néanmoins, un grand nombre de méthodes et critères permettant l'acquisition de la complexité. Par exemple, il existe plusieurs variations du critère GCV [32] ou encore des points de vue Bayesiens sur les splines de lissages permettant d'acquérir plus facilement ces paramètres [87, 136].

Un approfondissement de cette étude serait de questionner le modèle multiplicatif contre

l'alternative d'un modèle plus simple tel qu'un modèle additif, ou encore celle d'un modèle plus complexe, comprenant plus de composantes multiplicatives (c.f section 1.5). Par exemple, au travers de tests prenant en compte la dépendance temporelle (e.g. modèle ARMA), nous pourrions observer sur quelles localisations spatiales et sur quelles variables climatiques ce modèle contient toute l'information sur l'évolution du changement climatique. La modélisation proposée dans le chapitre 3 pour décrire l'espérance des variables de températures est bien sûr applicable à d'autres variables telles que l'aire et le volume de la banquise Arctique (voir Appendice pour plus de détails) ou toute autre variable d'intérêt : précipitation, intensité de vent, amplitude thermique journalière, entre autres. Il est possible d'utiliser ces normales pour décrire le changement climatique plus finement notamment via l'évolution de la classification climatique (des climats que disparaissent/apparaissent) basée sur une distance sur les normales estimées ou encore via les analogues climatiques. Autant pour les normales que pour les quantiles, la description offerte par ce point de vue est plus riche et précise que les précédentes méthodes et permettrait des applications multiples tant pour le suivi que pour l'adaptation climatique, comme les analogues développées par [33] ou encore re-visiter la classification climatique.

4.2.2 Analogues climatiques

Les normales estimées sur des scénarios de changement climatiques permettent de revisiter le calcul d'analogues climatiques. Il s'agit de rechercher pour un lieu x_0 au temps t_0 des lieux x_1 présentant un climat similaire au temps t_1. Nous avons calculé des analogues, pour le climat européen pré-industriel (1850) et le climat actuel, et ce, à l'aide d'une distance en norme L^2 entre les cycles annuels dont une représentation est illustrée dans la figure 4.15. Les analogues sont estimées à l'aide de données modèles. Les cartes suivantes indiquent les climats proches de celui de Toulouse en 2100 par des couleurs chaudes. Nous pouvons remarquer le rapprochement du climat futur de Toulouse avec des climats actuels ou passés est toujours situés à de plus faibles latitudes. Le meilleur analogue est atteint en Espagne proche de la frontière portugaise dans la ville de Badajoz.

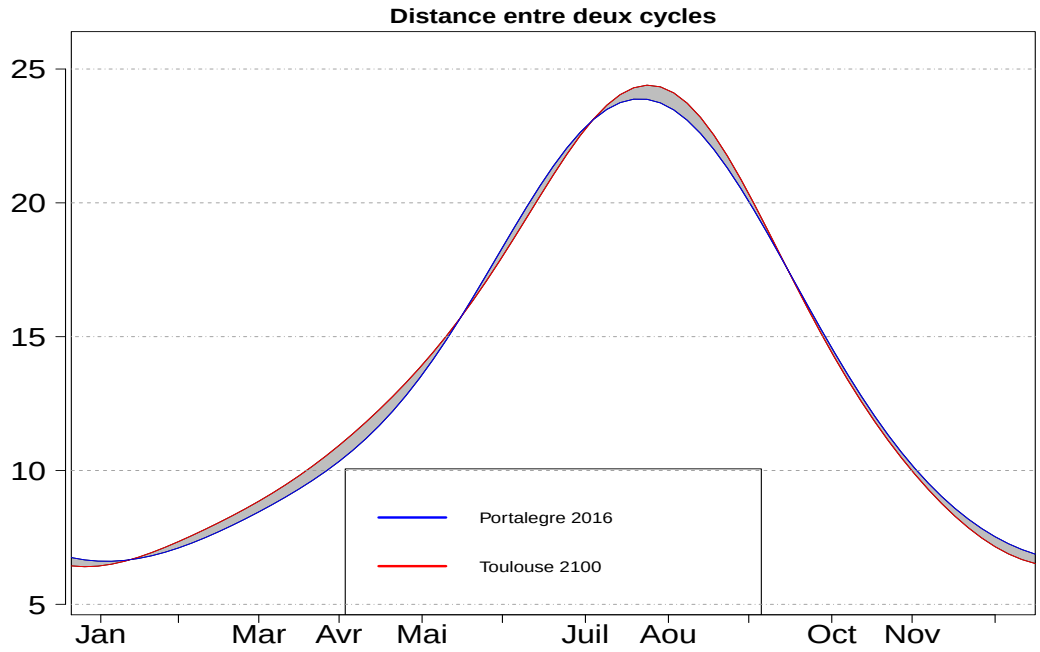


FIGURE 4.13 – Différence entre le cycle annuel de Portalegre en 2016 et de Toulouse en 2100.

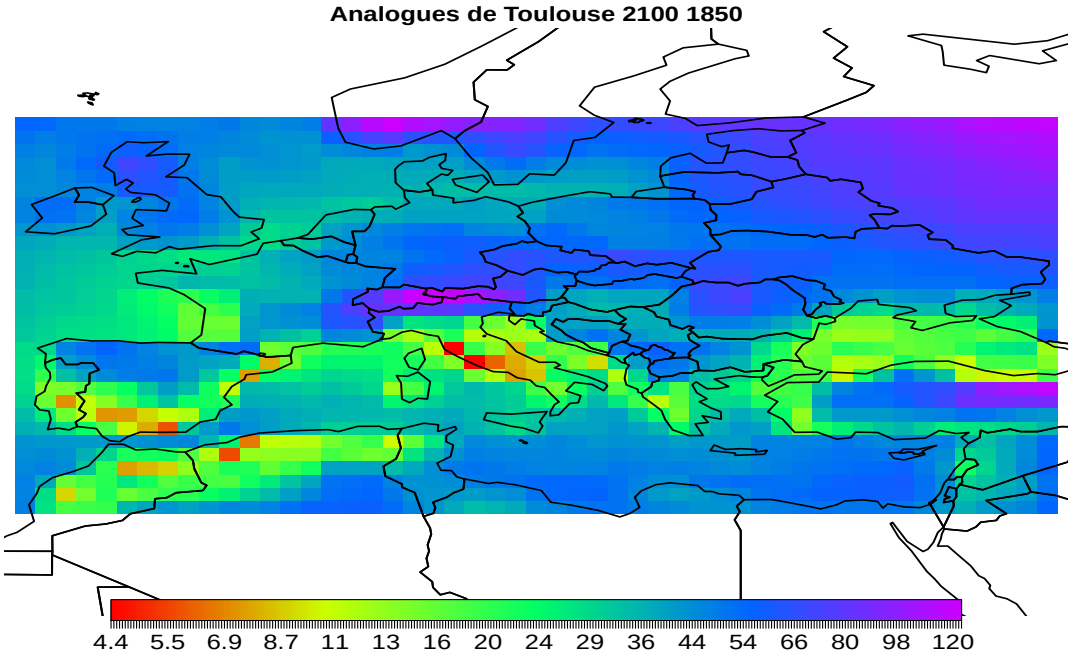


FIGURE 4.14 – Analogues en 1850 du climat futur de Toulouse en 2100.

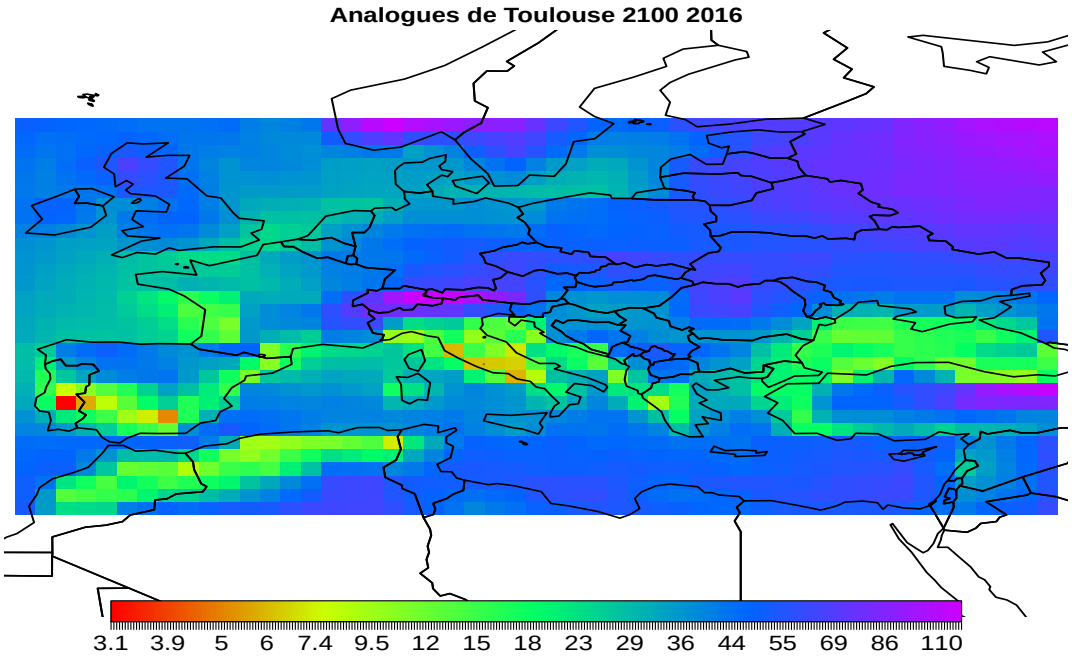


FIGURE 4.15 – Analogues actuelles (2016) du climat futur de Toulouse en 2100. Les couleurs représentent la distance à la normale de température de 2100.

Enfin, nos analogues climatiques avec un seul paramètre (la température) se rapprochent des précédents calculs d'analogues par exemple Kopf et al [82] (cf figure 1.7). Cela atteste de l'impact majeur des variables de températures sur les contraintes climatiques de Paris en 2100. Néanmoins, la caractérisation du climat ne peut, de manière générale, se faire sans l'intégration de plusieurs variables (cf figure 1.5). Il serait alors nécessaire de les prendre en compte au travers d'une distance adaptée sur des normales de plusieurs types de paramètres (températures et précipitations au minimum). Dans un premier temps, nous pourrions, par exemple, évaluer de la même façon que pour les températures, la distance en norme $\mathcal{L}^2$ (c.f figure 4.13) entre les normales de précipitations. Puis, dans un deuxième temps, définir une distance comme combinaison convexe de celles définies par les normales de températures et de précipitations. Pour obtenir des métriques encore plus fines, il est possible de reproduire la méthodologie précédente sur les distributions de probabilités estimées à l'aide de la régression quantile. Une métrique intéressante semble être donnée par la distance de Wasserstein.

Une application directe de ce type d'information serait, par exemple, de pouvoir situer un jour du climat actuel par rapport aux climats passés ou futurs au vue de la métrique choisie.

Une dernière idée serait de faire, à l'instar de la classification de Köppen développée dans la section 1.4.1, de la classification climatique. Ce qui permettrait d'observer l'apparition et disparition des climats au cours du temps.

4.2.3 Dynamique temporelle et générateurs de temps

Une première extension serait l'étude de la dépendance temporelle à courte échéance des températures moyennes. Un premier élément pour satisfaire cette étude serait de se placer dans un modèle Markovien. Nous nous placerons, dans un premier temps, dans le cadre simplifié d'une chaîne de Markov homogène (à temps et espace d'états discret) et où les lois marginales sont uniformes (figure 4.16). En effet, une fois en main une estimation de la fonction quantile et donc de la fonction de répartition, il est aisé de se ramener à une loi uniforme (lorsque la variable est continue). Ceci permet de se soustraire à la question des cycles saisonniers et des tendances dues au réchauffement et ainsi connaître la persistance moyenne relativement à la localisation des évènements dans leurs distributions. En outre, cette méthodologie permet d'estimer la probabilité de certains évènements extrêmes, tels que le nombre de jours au-delà du quantile .90.

Par soucis de continuité avec notre travail sur les quantiles, nous avons choisi d'obtenir l'intégralité de la loi dans ses valeurs les plus extrêmes à partir de l'estimation faite de ces derniers. Pour ce faire, nous avons simplement prolongé les quantiles par une loi de Pareto, dont le paramètre de forme $\gamma \neq 0$ est supposé non-nul, à l'aide de ses trois quantiles les plus extrêmes (i.e. les centiles 97, 98, 99 et 1, 2, 3). Allant dans le sens des travaux de [102] sur la médiane, cette méthodologie pourrait permettre d'avoir un estimateur plus robuste. Les quantiles centraux (0.03 et 0.97) servent alors de valeurs seuils, les quantiles restants servent à déterminer les paramètres de forme et de dispersion. Nous avons choisi la paramétrisation suivante pour les quantiles de la loi de Pareto [39][65] : $Q(\tau) = \frac{\sigma}{\gamma}(1 - (1 - \tau)^{\gamma})$. La figure A.2 (reportée en annexe) montre l'évolution des paramètres des lois de Pareto sur la période observée. Des méthodes semi-paramétriques [47] sembleraient être appropriées pour estimer les loi de Pareto.

Dans la figure 4.17 nous pouvons voir l'estimation de la matrice stochastique de la

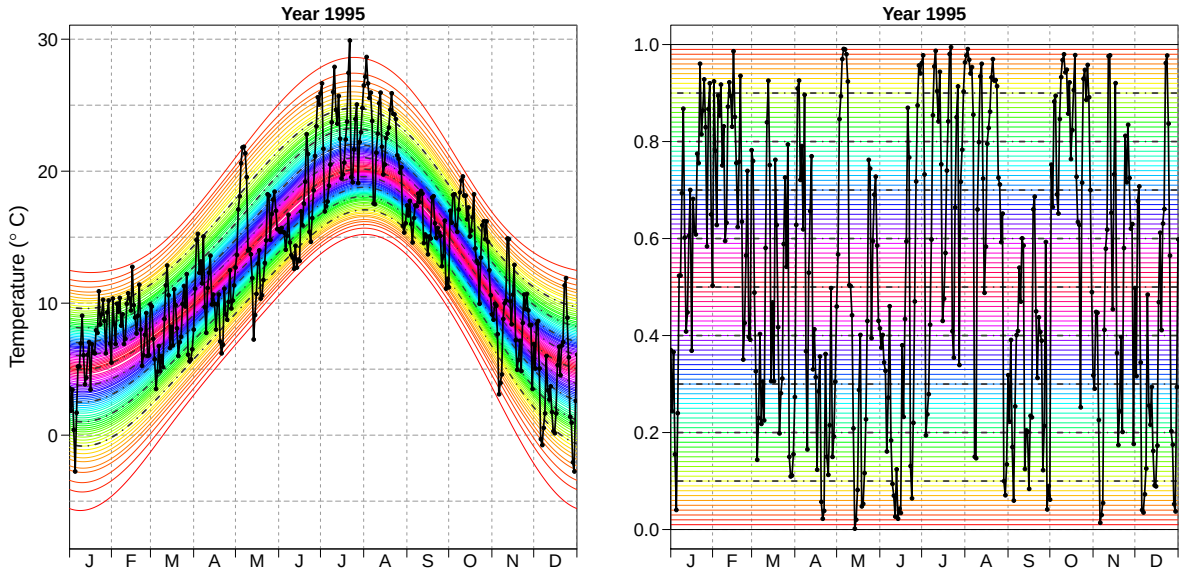


FIGURE 4.16 – Transformation du jeu de données corrigées de températures moyennes quotidiennes, sur Paris-Montsouris (par la fonction de répartition estimée par régression quantile) pour l'année 1995. Sur les deux figures, on peut voir la série chronologique en noir et les centiles représentés par les courbes en couleur et pointillées. La figure de droite est la transformation de la figure de gauche par leurs fonctions de répartition estimées pour chaque jour de l'année.

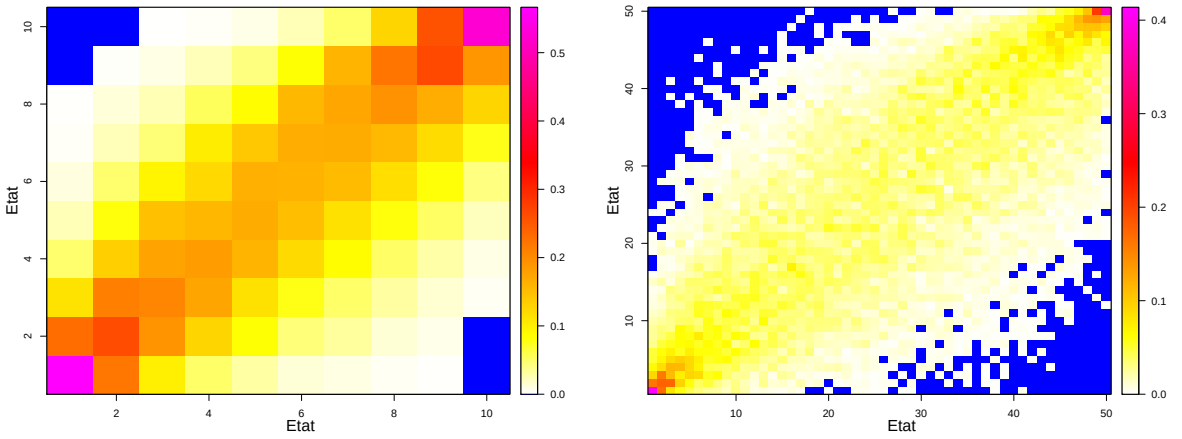


FIGURE 4.17 – Estimation de la matrice de transition avec 10 états à gauche et 50 états à droite, les états étant définis à l'aide des quantiles, plus exactement grâce aux déciles à gauche et aux centiles paires à droite. Les zones en bleu représentent des évènements ayant une probabilité nulle.

chaîne de Markov pour des états définis par les intervalles dont les bornes sont les quantiles consécutifs considérés $]Q_1, Q_2]$. Autrement dit, chaque ligne i de la matrice représente la probabilité d'être dans l'état j (représenté par les colonnes) le jour suivant. Plus pré-

cisément, dans le panel de gauche, on considère les déciles ; les états sont définis par les intervalles $]-\infty, Q_{0.1}]$, $]Q_{\frac{i}{10}}, Q_{\frac{i}{10}+0.1}]$, pour $i \in [\![1, 8]\!]$, $]Q_{0.9}, \infty]$ et dans le panel de droite, ils sont définis par les centiles paires $]-\infty, Q_{0.02}]$, $]Q_{\frac{2i}{100}}, Q_{\frac{2i+2}{100}+0.1}]$, pour $i \in [\![1, 49]\!]$, $]Q_{0.98}, \infty]$. Cette estimation est beaucoup plus bruitée que la précédente. En effet, pour obtenir la probabilité d'appartenance à un état, conditionnellement à une classe fixée (loi multinomiale), nous ne disposons plus que de 430 observations contre 2153 pour la matrice stochastique de gauche.

Quel que soit l'état considéré, l'état suivant a une forte probabilité d'appartenir aux états les plus proches. Cette loi de probabilité tend à être plus diffuse sur les états centraux et plus concentrée sur les parties les plus extrêmes de la distribution. Par exemple, les classes extrêmes ont une probabilité supérieure à 0.6 de rester dans leurs états actuels ou états adjacents le jour suivant. Ce phénomène est très localisé au-dessus du $95^{ème}$ centile et en-dessous du $5^{ème}$ centile comme le montre le panel de droite de la figure 4.17. De plus, nous estimons qu'il y a une probabilité nulle de passer d'un extrême chaud à un extrême froid (signifiant qu'aucun cas de ce type n'a été observé sur la période considérée de 1959-2017). Globalement, la dépendance temporelle ressemble beaucoup à celle d'un processus gaussien auto-régressif, cependant le modèle Markovien montre une légère dissymétrie entre les extrêmes chauds et froids. On peut alors améliorer ce modèle en lui permettant d'avoir une évolution temporelle à l'aide d'un modèle de régression logistique multinomiale.

Cette description reste très simpliste. En effet, les évènements considérés sont de nature discrète et, comme le montre la figure précédente, ne nous permettent pas d'avoir une description continue de la dépendance conditionnelle. À cet égard, des méthodes telles que la théorie des copules, pourvoiraient une meilleure approche ou encore des modèles de type (régression quantile) QAR (cf section 2.2.6). Cependant, la prise en compte d'une tendance dans ces modèles semble compliquée à obtenir. De surcroît, la dépendance pourrait impliquer une profondeur temporelle plus grande, par exemple, via des chaînes de Markov d'ordre supérieur à 1, rendant le modèle précédent impraticable sur des séries d'observations sans hypothèses plus contraintes. Par exemple, avec une chaîne de Markov d'ordre de 2, nous serions amenés à considérer 100 états, ce qui produirait une estimation encore plus bruitée que la figure 4.17. Plus dérangeant encore, la dépendance pourrait impliquer des variables éloignées

dans le temps (*long-range dependance*), par exemple, dépendant de la température moyenne décennale.

Ce type de dépendance devrait pouvoir être modélisée avec des chaînes de Markov cachées dont les états sous-jacents ne sont pas observés. Le modèle devient alors plus malléable mais aussi moins interprétable car nous n'avons, à posteriori, pas d'information directe sur ces "états cachés".

4.2.4 Distributions multivariées

Ces méthodes peuvent s'étendre au cadre multivarié, par exemple, en considérant les travaux de [3] sur les surfaces quantiles. Ces derniers sont un des prolongements naturels des quantiles 1D aux quantiles multivariés et permettent un prolongement immédiat des modèles de régressions quantiles proposés. Les "surfaces quantiles" permettent notamment de caractériser une loi en dimension supérieure.

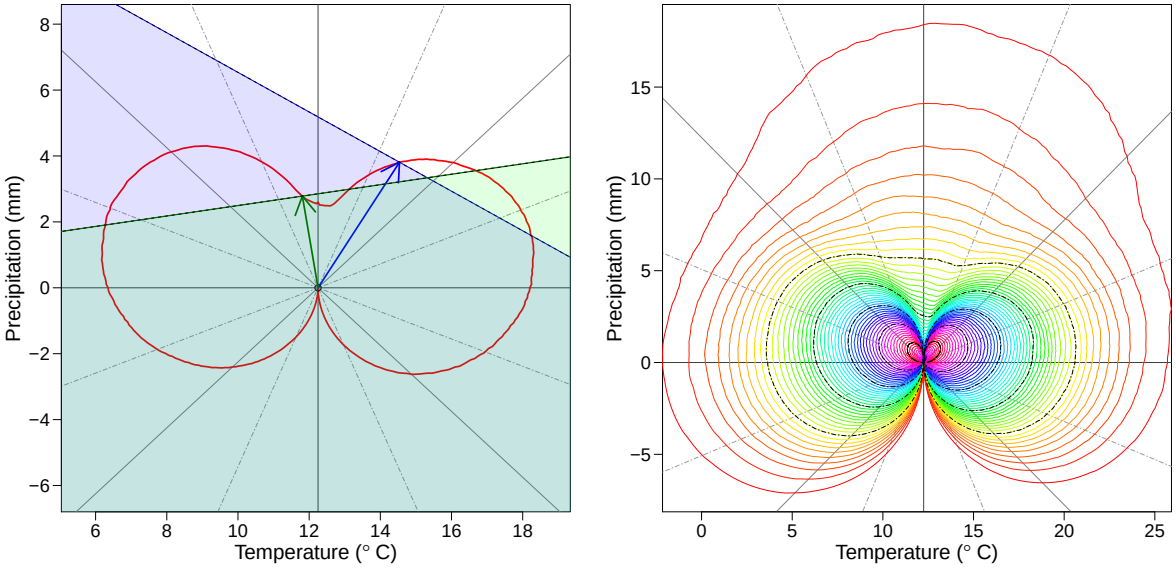


FIGURE 4.18 – La figure de gauche montre la construction des courbes quantiles à partir des demi-espaces. La figure de droite montre l'ensemble de ses courbes pour les centiles.

La figure 4.18 montre dans le panel de gauche la construction des courbes quantiles (en rouge) pour $\alpha = 0.8$ vu du point O. Le point O est positionné sur les médianes respectives de deux variables, 12.25 pour les températures, 0 pour les précipitations. Les deux demi-plans

(bleu et vert) correspondent alors à des zones de probabilité α telles que leurs frontières soient orthogonales à leurs directions d'observation respectives, représentées par les deux vecteurs partant de l'origine O. Le panel de droite montre l'ensemble de ces courbes pour des valeurs de $\alpha \in \{\frac{50+i}{100}, \quad i \in [\![0, 49]\!]\}$. Pour ajouter une tendance à cette construction, il suffirait d'appliquer un modèle de régression quantile en chacune des directions. Ceci demanderait, néanmoins, une sélection de la complexité là aussi pour chacune de ces directions. Nous aurions alors caractérisé l'évolution de la loi (2D) sous l'effet du changement climatique. Il y a, bien sûr, d'autres approches pour appréhender les quantiles dans le cas multivarié ; on peut citer [52] ou encore utiliser la théorie des copules.

Ici aussi nous pourrions définir une métrique entre l'évolution saisonnière des distributions de variables climatiques, par exemple, via la distance Wasserstein 1D ou encore à l'aide des surfaces quantiles et d'une distance appropriée.

Annexe A

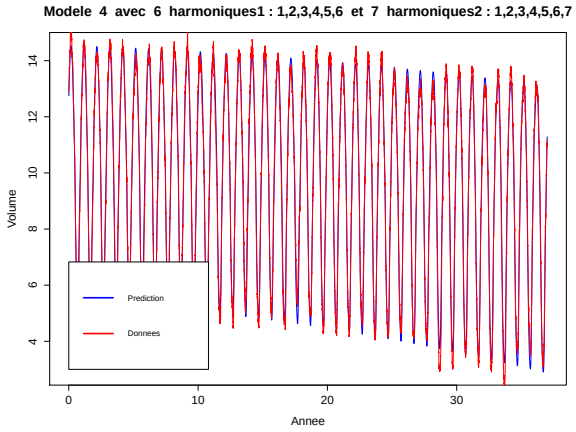
Appendice

A.1 Aire et volume de banquise Arctique

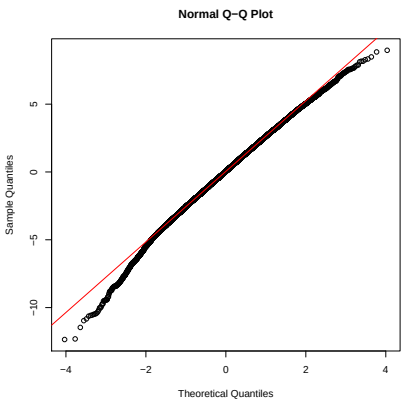
Les données issues du modèle PIOMAS (Pan-Artic Ice Ocean Modelling and Assimilation System) sont constituées d'aires et de volumes de banquise Arctique au pas de temps journalier du 1_{er} janvier 1979 au 20 mars 2016, soit 37 ans et trois mois. Le volume est exprimé en milliers de km^3 et l'aire en millions de km^2. Nous pouvons étudier l'évolution de la banquise avec les modèles de régression non-linéaire développés dans le chapitre 3.

Aire Pour les données d'aires c'est le modèle 3 (section 3.2.1.2) sélectionné par BIC qui offre les meilleurs résultats. L'erreur quadratique moyenne (train) est de 0.127 et la variance expliquée de 0.99. Ceci montre un excellent ajustement aux données et un bruit très faible. Le nombre d'harmoniques utilisées pour décrire le cycle annuel de référence est de 6, et de 7 pour le delta cycle.

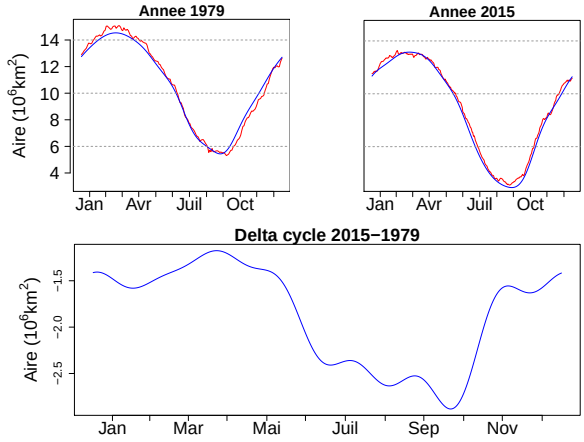
Prédiction



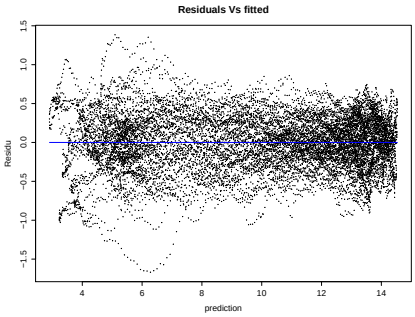
QQplot



Cycles et Delta Cycles

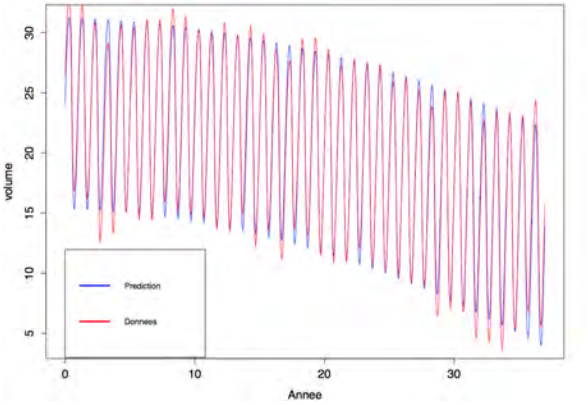


Résidus

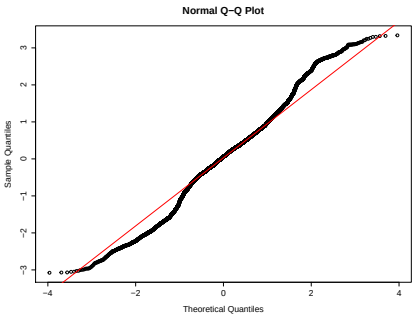


Volume De même que pour les données d'aire, le volume de banquise est mieux décrit par le modèle 3 (section 3.2.1.2). Tout comme les scores précédents, les données de volume de banquise bénéficient d'un excellent ratio signal sur bruit, nous obtenons un coefficient de détermination R^2 de 0.97 et une erreur quadratique moyenne de 1.209.

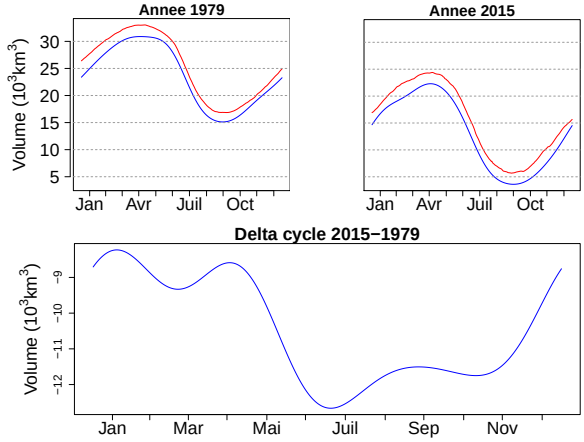
Prédiction



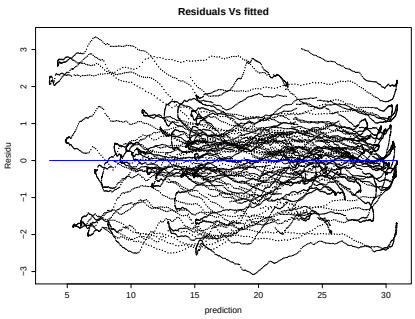
QQplot



Cycles et Delta Cycles



Résidus



A.2 Minimales et maximales de températures sur un échantillon divisé par 15

Tout comme pour les températures moyennes, dans le cas des minimales et maximales de température, le modèle de Krigeage donnant les meilleurs résultats est le Krigeage avec fonction de covariance Matèrn $\frac{3}{2}$. Les modèles 2 et 3 ont alors les mêmes performances en terme d'erreur quadratique et procure des résultats similaires. Nous montrons, dans les figures suivantes, les estimations faites par le modèle de Krigeage sur les deux variables.

Krigeage avec covariance Matérn $\frac{3}{2}$

Remarque A.1

On peut s'étonner de la différence de comportement des minimales $Tmin$, moyennes $Tmoy$ et maximales $Tmax$ notamment en terme d'erreur quadratique moyenne.
Cependant, en estimant la variabilité de ces variables ainsi que leurs covariances, nous obtenons :
$\widehat{var}(Tmoy) = 6.88$ $\widehat{var}(Tmin) = 6.95$ $\widehat{var}(Tmax) = 9.71$ $\widehat{cov}(Tmax, Tmin) = 5.49$
De plus, en ayant en tête que pour des raisons historiques $Tmoy = \frac{Tmin-Tmax}{2}$, il vient :
$var(Tmoy) = \frac{1}{4}(var(Tmax) + 2.cov(Tmax, Tmin) + var(Tmin))$
ce qui semble aller dans le sens des estimations précédentes car :
$\frac{1}{4}(\widehat{var}(Tmax) + 2.\widehat{cov}(Tmax, Tmin) + \widehat{var}(Tmin)) = 6.91$

A.3 Évolution des paramètres de la loi de Pareto

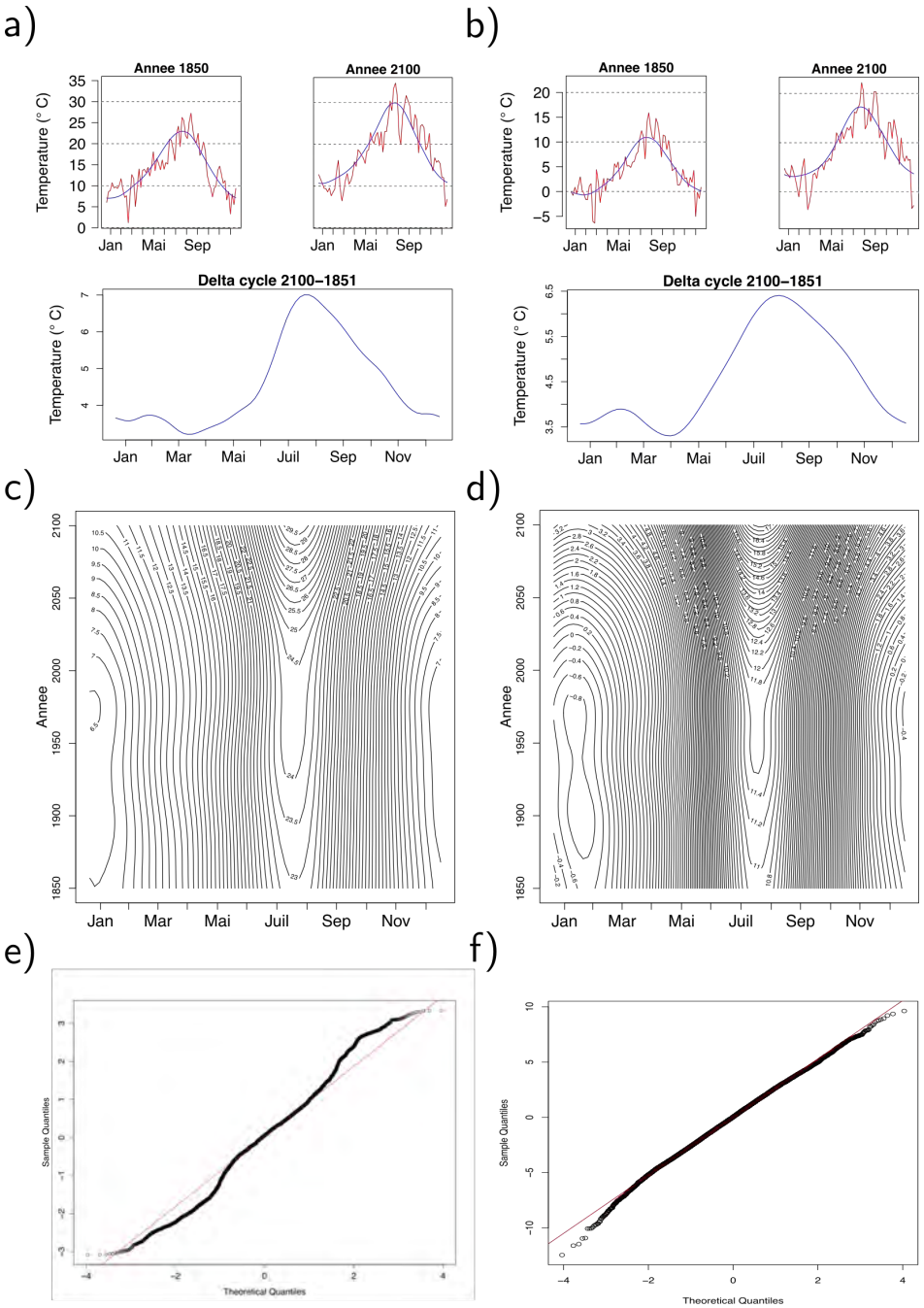


FIGURE A.1 – La première colonne (colonne de gauche) montre les résultats associés aux maximales de température, la colonne de droite ceux associés aux minimales de températures, les deux estimations sont issues d'un Krigeage avec covariance Matèrne 3/2.

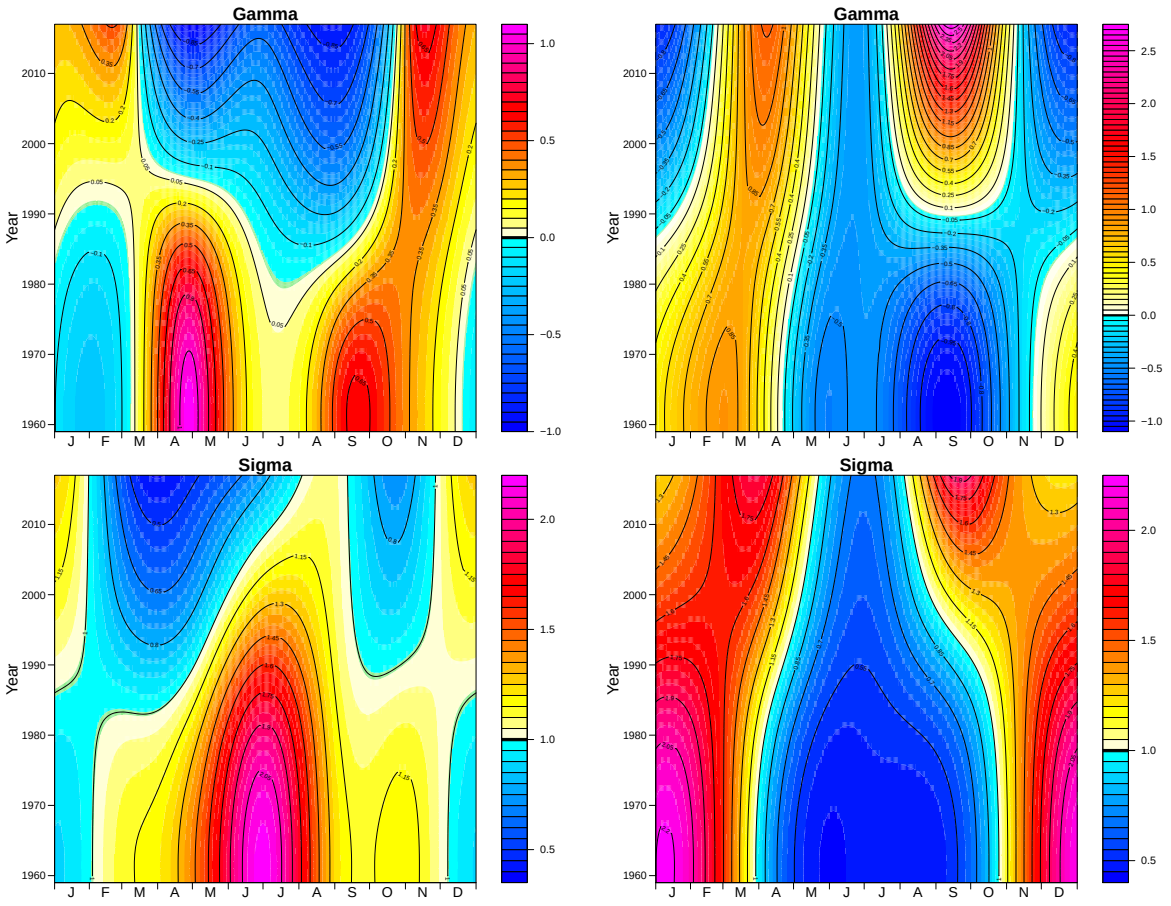


FIGURE A.2 – Évolution des paramètres de la loi de Pareto à partir des trois derniers (resp. premiers) quantiles estimés par régression quantile (modèle 4). Les paramètres pour les extrêmes chauds de températures sont décrits par les panels de gauche, les extrêmes froids par les panels de droite.

Bibliographie

[1] Milton Abramowitz et Irene A Stegun. *Handbook of mathematical functions : with formulas, graphs, and mathematical tables*. T. 55. Courier Corporation, 1965 (cf. p. 39).

[2] NN Adelmalek. "On the discrete linear L1 approximation and L1 solutions of overdetermined linear equations". In : *J. Approximation Theory* 11 (1974), p. 38–53 (cf. p. 46).

[3] Adil Ahidar-Coutrix. "Surfaces quantile : propriétés, convergences et applications". Thèse de doctorat dirigée par Berthet, Philippe et Klein, Thierry Mathématiques appliquées Toulouse 3 2015. Thèse de doct. 2015 (cf. p. 177).

[4] H. Akaike. "A new look at the statistical model identification". In : *IEEE Transactions on Automatic Control* 19.6 (1974), p. 716–723 (cf. p. 64).

[5] Lisa V Alexander. "Global observed long-term changes in temperature and precipitation extremes : a review of progress and limitations in IPCC assessments and beyond". In : *Weather and Climate Extremes* 11 (2016), p. 4–16 (cf. p. 147).

[6] Tsuneo Arakawa et al. *Bernoulli numbers and zeta functions*. Springer, 2014 (cf. p. 39).

[7] Aleksandr Y Aravkin et al. *The connection between Bayesian estimation of a Gaussian random field and RKHS, Submitted to IEEE Transactions on Neural Networks and Learning Systems* (cf. p. 111).

[8] Anthony Arguez et Scott Applequist. "A Harmonic Approach for Calculating Daily Temperature Normals Constrained by Homogenized Monthly Temperature Normals". In : *Journal of Atmospheric and Oceanic Technology* 30.7 (2013), p. 1259–1265. eprint : `https://doi.org/10.1175/JTECH-D-12-00195.1` (cf. p. 75, 79).

[9] Anthony Arguez et Russell S. Vose. "The definition of the standard WMO climate normal : The key to deriving alternative climate normals". In : *Bulletin of the American Meteorological Society* 92.6 (2011), p. 699–704 (cf. p. 74).

[10] Anthony Arguez, Russell S. Vose et Jenny Dissen. “Alternative climate normals : Impacts to the energy industry”. In : *Bulletin of the American Meteorological Society*. T. 94. 6. 2013, p. 915–917 (cf. p. 75).

[11] Anthony Arguez et al. “Noaa’s 1981-2010 U.S. climate normals”. In : *Bulletin of the American Meteorological Society* 93.11 (2012), p. 1687–1697 (cf. p. 75).

[12] Anthony Arguez et al. *NOAA’s 1981–2010 climate normals : Methodology of temperature-related normals*. Rapp. tech. 2011, p. 7 (cf. p. 75, 128, 169).

[13] Sylvain Arlot, Alain Celisse et al. “A survey of cross-validation procedures for model selection”. In : *Statistics surveys* 4 (2010), p. 40–79 (cf. p. 67).

[14] Rebecca G Asch et al. “Demystifying models : answers to ten common questions that ecologists have about Earth system models”. In : (2016) (cf. p. 12).

[15] Jean Marc Azaïs et Aurélien Ribes. “Multivariate spline analysis for multiplicative models : Estimation, testing and application to climate change”. In : *Journal of Multivariate Analysis* 144 (2016), p. 38–53 (cf. p. 81, 107, 108, 135, 140).

[16] Omar Baddour. “Climate Normals, World Meteorological Organization Commission for Climatology Management Group Meeting , ITEM 10 ”. In : (2011) (cf. p. 75).

[17] Eve Bofinger. “NON-PARAMETRIC ESTIMATION OF DENSITY FOR REGULARLY VARYING DISTRIBUTIONS1”. In : *Australian Journal of Statistics* 17.3 (1975), p. 192–195 (cf. p. 52).

[18] Aurélie Boisbunon et al. “AIC, Cp and estimators of loss for elliptically symmetric distributions”. In : *arXiv preprint arXiv :1308.2766* (2013) (cf. p. 66).

[19] Bertrand Bonan. “Assimilation de données pour l’initialisation et l’estimation de paramètres d’un modèle d’évolution de calotte polaire”. Thèse de doct. Université de Grenoble, 2013 (cf. p. 14).

[20] Ronald J Bosch, Yinyu Ye et George G Woodworth. “A convergent algorithm for quantile regression with smoothing splines”. In : *Computational statistics & data analysis* 19.6 (1995), p. 613–630 (cf. p. 56, 138).

[21] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010 (cf. p. 35).

[22] T. A. Buishand, M. V. Shabalova et T. Brandsma. “On the Choice of the Temporal Aggregation Level for Statistical Downscaling of Precipitation”. In : *Journal of Climate* 17.9 (2004), p. 1816–1827. eprint : https://doi.org/10.1175/1520-0442(2004)017<1816:OTCOTT>2.0.CO;2 (cf. p. 137).

[23] Prabir Burman. “A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods”. In : *Biometrika* 76.3 (1989), p. 503–514 (cf. p. 67).

[24] Julien Cattiaux et Aurélien Ribes. “Defining single extreme weather events in a climate perspective”. In : *Bulletin of the American Meteorological Society* 99.8 (2018), p. 1557–1568 (cf. p. 133).

[25] Joseph E Cavanaugh et Andrew A Neath. “Akaike’s information criterion : Background, derivation, properties, and refinements”. In : *International Encyclopedia of Statistical Science* (2011), p. 26–29 (cf. p. 66).

[26] JM Chambers et al. “Graphical methods for data analysis, ser”. In : *The Wadsworth statistics/probability series. Wadsworth & Brooks/Cole, Pacific Grove, CA* (1983) (cf. p. 144, 150, 165).

[27] Victor Chernozhukov, Iván Fernández-Val et Alfred Galichon. “Quantile and probability curves without crossing”. In : *Econometrica* 78.3 (2010), p. 1093–1125 (cf. p. 56, 135, 141).

[28] Victor Chernozhukov et Han Hong. “Three-step censored quantile regression and extramarital affairs”. In : *Journal of the American Statistical Association* 97.459 (2002), p. 872–882 (cf. p. 45).

[29] Victor Chernozhukov et al. “Extremal quantile regression”. In : *The Annals of Statistics* 33.2 (2005), p. 806–839 (cf. p. 155).

[30] National Research Council et al. *Assessment of intraseasonal to interannual climate prediction and predictability*. National Academies Press, 2010 (cf. p. 91).

[31] P Craven. “G Wahba Smoothing noisy data with spline functions”. In : *Numerische Mathematik* 31 (1979), p. 377–403 (cf. p. 38).

[32] Peter Craven et Grace Wahba. “Smoothing noisy data with spline functions”. In : *Numerische mathematik* 31.4 (1978), p. 377–403 (cf. p. 170).

[33] Fabienne Dahinden, Erich M Fischer et Reto Knutti. “Future local climate unlike currently observed anywhere”. In : *Environmental Research Letters* 12.8 (2017), p. 084004 (cf. p. 171).

[34] Holger Dette et Stanislav Volgushev. “Non-crossing non-parametric estimates of quantile curves”. In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 70.3 (2008), p. 609–627 (cf. p. 57).

[35] Pedro Domingos. “A unified bias-variance decomposition for zero-one and squared loss”. In : *AAAI/IAAI* 2000 (2000), p. 564–569 (cf. p. 62).

[36] Francis Ysidro Edgeworth. “XXII. On a new method of reducing observations relating to several quantities”. In : *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 25.154 (1888), p. 184–191 (cf. p. 40).

[37] B Efron. “Bootstrap Methods : Another Look at the Jackknife. Ann. Stat”. In : (1979) (cf. p. 52).

[38] Faouzi El Bantli et Marc Hallin. “L1-estimation in linear models with heterogeneous white noise”. In : *Statistics & probability letters* 45.4 (1999), p. 305–315 (cf. p. 47).

[39] Soheil Saeed Far et Ahmad Khairi Abd Wahab. “Evaluation of peaks-over-threshold method”. In : *Ocean Science Discussions* (2016), p. 1–25 (cf. p. 174).

[40] Jean Favard. “Sur l’interpolation”. In : *Bulletin de la Société Mathématique de France* 67 (1939), p. 102–113 (cf. p. 162).

[41] Martin Fox, Herman Rubin et al. “Admissibility of quantile estimates of a single location parameter”. In : *The Annals of Mathematical Statistics* 35.3 (1964), p. 1019–1030 (cf. p. 42).

[42] Jerome Friedman, Trevor Hastie et Robert Tibshirani. *The elements of statistical learning*. T. 1. 10. Springer series in statistics New York, NY, USA : 2001 (cf. p. 24, 35).

[43] Tadayoshi Fushiki. “Estimation of prediction error by using K-fold cross-validation”. In : *Statistics and Computing* 21.2 (2011), p. 137–146 (cf. p. 67).

[44] Carl Friedrich Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. T. 7. Perthes et Besser, 1809 (cf. p. 40).

[45] Rudolf Geiger. “Klassifikation der klimate nach W. Köppen”. In : *Landolt-Börnstein–Zahlenwerte und Funktionen aus Physik, Chemie, Astronomie, Geophysik und Technik* 3 (1954), p. 603–607 (cf. p. 9).

[46] Olivier Geoffroy et David Saint-Martin. “Pattern decomposition of the transient climate response”. In : 66 (mar. 2014), p. 23393 (cf. p. 81, 141).

[47] Irène Gijbels, Rezaul Karim et Anneleen Verhasselt. “Quantile estimation in a generalized asymmetric distributional setting”. In : *Workshop on Stochastic Models, Statistics and their Application*. Springer. 2019, p. 13–40 (cf. p. 174).

[48] David Ginsbourger. “Multiples métamodèles pour l’approximation et l’optimisation de fonctions numériques multivariables”. Thèse de doctorat dirigée par Carraro, Laurent Mathématiques appliquées Saint-Etienne, EMSE 2009. Thèse de doct. 2009, 1 vol. (375 p.) (Cf. p. 26).

[49] Federico Girosi, Michael Jones et Tomaso Poggio. “Regularization theory and neural networks architectures”. In : *Neural computation* 7.2 (1995), p. 219–269 (cf. p. 36).

[50] Pauline Givord, Xavier Dhaultfoeuille et al. *La régression quantile en pratique*. Rapp. tech. Institut National de la Statistique et des Etudes Economiques, 2013 (cf. p. 47).

[51] Peter Hall et Simon J Sheather. “On the distribution of a studentized quantile”. In : *Journal of the Royal Statistical Society : Series B (Methodological)* 50.3 (1988), p. 381–391 (cf. p. 52).

[52] Marc Hallin et al. “On distribution and quantile functions, ranks and signs in R d”. In : *ECARES Working Papers* (2017) (cf. p. 178).

[53] Trevor Hastie, Robert Tibshirani et Martin Wainwright. *Statistical learning with sparsity : the lasso and generalizations*. CRC press, 2015 (cf. p. 65, 86).

[54] Cecil Hastings et al. “Low moments for small samples : a comparative study of order statistics”. In : *The Annals of Mathematical Statistics* 18.3 (1947), p. 413–426 (cf. p. 40).

[55] Matz A Haugen et al. “Estimating Changes in Temperature Distributions in a Large Ensemble of Climate Simulations Using Quantile Regression”. In : *Journal of Climate* 31.20 (2018), p. 8573–8588 (cf. p. 133, 134, 139, 169).

[56] MR Haylock et al. “A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006”. In : *Journal of Geophysical Research : Atmospheres* 113.D20 (2008) (cf. p. 125).

[57] Xuming He et Feifang Hu. “Markov chain marginal bootstrap”. In : *Journal of the American Statistical Association* 97.459 (2002), p. 783–795 (cf. p. 54).

[58] Andrew John Herbertson. *Outlines of physiography : An introduction to the study of the earth*. Arnold, 1908 (cf. p. 4).

[59] *Historical Overview of Climate Change Science. Chapter 1* (cf. p. 12).

[60] Nils Lid Hjort et David Pollard. “Asymptotics for minimisers of convex processes”. In : *arXiv preprint arXiv :1107.3806* (2011) (cf. p. 50).

[61] Reza Hosseini. “Quantiles equivariance”. In : *arXiv preprint arXiv :1004.0533* (2010) (cf. p. 44).

[62] Jin Huang, Huug M van den Dool et Anthony G Barnston. “Long-lead seasonal temperature prediction using optimal climate normals”. In : *Journal of Climate* 9.4 (1996), p. 809–817 (cf. p. 78).

[63] Gareth M James. “Variance and bias for general loss functions”. In : *Machine learning* 51.2 (2003), p. 115–135 (cf. p. 62).

[64] Catherine Jeandel et Rémy Mosseri. “Le climat à découvert”. In : *Paris : CNRS Editions* (2011) (cf. p. 13).

[65] Jelena Jocković. “Quantile estimation for the generalized pareto distribution with application to finance”. In : *Yugoslav Journal of Operations Research* 22.2 (2016) (cf. p. 174).

[66] Dikra Khedhaouiria, Alain Mailhot et Anne-Catherine Favre. “Daily Precipitation Fields Modeling across the Great Lakes Region (Canada) by Using the CFSR Reanalysis”. In : *Journal of Applied Meteorology and Climatology* 57.10 (2018), p. 2419–2438 (cf. p. 135, 137).

[67] George Kimeldorf et Grace Wahba. “Some results on Tchebycheffian spline functions”. In : *Journal of mathematical analysis and applications* 33.1 (1971), p. 82–95 (cf. p. 30).

[68] Erik Kjellström et al. “European climate change at global mean temperature increases of 1.5 and 2 degrees C above pre-industrial conditions as simulated by the EURO-CORDEX regional climate models”. In : *Earth System Dynamics* 9.2 (2018), p. 459–478 (cf. p. 126).

[69] Keith Knight. “Limiting distributions for L1 regression estimators under general conditions”. In : *Annals of statistics* (1998), p. 755–770 (cf. p. 49).

[70] Masha Kocherginsky et Xuming He. “Extensions of the Markov chain marginal bootstrap”. In : *Statistics & Probability Letters* 77.12 (2007), p. 1258–1268 (cf. p. 54).

[71] Masha Kocherginsky, Xuming He et Yunming Mu. “Practical Confidence Intervals for Regression Quantiles”. In : *Journal of Computational and Graphical Statistics* 14.1 (2005), p. 41–55. eprint : `https://doi.org/10.1198/106186005X27563` (cf. p. 51).

[72] Masha Kocherginsky, Xuming He et Yunming Mu. “Practical confidence intervals for regression quantiles”. In : *Journal of Computational and Graphical Statistics* 14.1 (2005), p. 41–55 (cf. p. 54).

[73] R. Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005 (cf. p. 40, 42, 43, 47, 49, 51).

[74] Roger Koenker. “Additive models for quantile regression : Model selection and confidence bandaids”. In : *Brazilian Journal of Probability and Statistics* 25 (nov. 2011), p. 239–262 (cf. p. 138).

[75] Roger Koenker. “Confidence intervals for regression quantiles”. In : *Asymptotic statistics*. Springer, 1994, p. 349–359 (cf. p. 52, 53).

[76] Roger KOENKER. “Quantile regression : 40 years on”. In : *Annual Review of Economics* 9 (2017), p. 155–176 (cf. p. 40, 42, 55).

[77] Roger KOENKER et Gilbert BASSETT JR. “Regression quantiles”. In : *Econometrica : journal of the Econometric Society* (1978), p. 33–50 (cf. p. 40, 43, 46, 47).

[78] Roger KOENKER et Ivan MIZERA. “Penalized triograms : total variation regularization for bivariate smoothing”. In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 66.1 (2004), p. 145–163. eprint : `https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2004.00437.x` (cf. p. 138).

[79] ROGER KOENKER, PIN NG et STEPHEN PORTNOY. “Quantile smoothing splines”. In : *Biometrika* 81.4 (1994), p. 673–680. eprint : `/oup/backfile/content_public/journal/biomet/81/4/10.1093/biomet/81.4.673/2/81-4-673.pdf` (cf. p. 56, 162).

[80] Roger KOENKER et Beum J PARK. “An interior point algorithm for nonlinear quantile regression”. In : *Journal of Econometrics* 71.1-2 (1996), p. 265–283 (cf. p. 43).

[81] Roger KOENKER et Zhijie XIAO. “Quantile autoregression”. In : *Journal of the American Statistical Association* 101.475 (2006), p. 980–990 (cf. p. 55).

[82] Sebastian KOPF, Minh HA-DUONG et Stéphane HALLEGATTE. “Using maps of city analogues to display and interpret climate change scenarios and their uncertainty”. In : *Natural hazards and earth system sciences* 8 (2008), p. 905–918 (cf. p. 10, 173).

[83] Wladimir KÖPPEN, Esther VOLKEN et Stefan BRÖNNIMANN. “The thermal zones of the earth according to the duration of hot, moderate and cold periods and to the impact of heat on the organic world (Translated from : Die Wärmezonen der Erde, nach der Dauer der heissen, gemässigten und kalten Zeit und nach der Wirkung der Wärme auf die organische Welt betrachtet, Meteorol Z 1884, 1, 215-226)”. In : *Meteorologische Zeitschrift* 20.3 (2011), p. 351–360 (cf. p. 8).

[84] Nir Y. KRAKAUER. “Estimating climate trends : Application to united states plant hardiness zones”. In : *Advances in Meteorology* (2012) (cf. p. 80).

[85] Nir Y KRAKAUER et Naresh DEVINENI. “Up-to-date probabilistic temperature climatologies”. In : *Environ. Res. Lett. Environ. Res. Lett* 10.10 (2015) (cf. p. 75).

[86] Daniel G Krige. “A statistical approach to some basic mine valuation problems on the Witwatersrand”. In : *Journal of the Southern African Institute of Mining and Metallurgy* 52.6 (1951), p. 119–139 (cf. p. 110).

[87] Tatyana Krivobokova et Göran Kauermann. “A note on penalized spline smoothing with correlated errors”. In : *Journal of the American Statistical Association* 102.480 (2007), p. 1328–1337 (cf. p. 170).

[88] Hans R Kunsch. “The jackknife and the bootstrap for general stationary observations”. In : *The annals of Statistics* (1989), p. 1217–1241 (cf. p. 54).

[89] S D Attri A K Jaswal L S Rathore. “STATE LEVEL CLIMATE CHANGE TRENDS IN INDIA ”. In : (2013) (cf. p. 103).

[90] G Lefebvre. *Comparison of Meteorological Screens for Temperature Measurement, paper presented at TECO-98 (Casablanca)*. Rapp. tech. Instruments et Observing Methods Report, 1998 (cf. p. 21).

[91] Adrien Marie Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805 (cf. p. 40).

[92] Robert E. Livezey et Marina M. Timofeyeva. “The first decade of long-Lead U.S. seasonal forecasts”. In : *Bulletin of the American Meteorological Society* (2008) (cf. p. 128, 169).

[93] Robert E. Livezey et al. “Estimation and Extrapolation of Climate Normals and Climatic Trends”. In : *Journal of Applied Meteorology and Climatology* 46.11 (2007), p. 1759–1776. eprint : `https://doi.org/10.1175/2007JAMC1666.1` (cf. p. 75, 78, 79).

[94] C. L. Mallows. “Some Comments on CP”. In : *Technometrics* 15.4 (1973), p. 661–675 (cf. p. 64).

[95] Georges Matheron. “Principles of geostatistics”. In : *Economic geology* 58.8 (1963), p. 1246–1266 (cf. p. 110).

[96] Timothy D Mitchell. “An Examination of the Accuracy of the Technique for Describing Future Climates”. In : *Climatic Change* Volume 60.Issue 3 (2003), pp 217–242 (cf. p. 15, 81, 141).

[97] Ivan Mizera et Jon A Wellner. “Necessary and sufficient conditions for weak consistency of the median of independent but not identically distributed random variables”. In : *Annals of statistics* (1998), p. 672–691 (cf. p. 48).

[98] DS Moore. “Limiting distributions for sample quantiles”. In : *The American Mathematical Monthly* 76.8 (1969), p. 927–929 (cf. p. 43).

[99] LR Mudryk, PJ Kushner et Chris Derksen. “Interpreting observed Northern Hemisphere snow trends with large ensembles of climate simulations”. In : *Climate dynamics* 43.1-2 (2014), p. 345–359 (cf. p. 125).

[100] Andrew A Neath et Joseph E Cavanaugh. “The Bayesian information criterion : background, derivation, and applications”. In : *Wiley Interdisciplinary Reviews : Computational Statistics* 4.2 (2012), p. 199–203 (cf. p. 66).

[101] MI Parzen, LJ Wei et Z Ying. “A resampling method based on pivotal estimating functions”. In : *Biometrika* 81.2 (1994), p. 341–350 (cf. p. 53).

[102] Liang Peng et AH Welsh. “Robust estimation of the generalized pareto distribution”. In : *Extremes* 4.1 (2001), p. 53–65 (cf. p. 174).

[103] David Pollard. “Asymptotics for least absolute deviation regression estimators”. In : *Econometric Theory* 7.2 (1991), p. 186–199 (cf. p. 49).

[104] Stephen Portnoy, Roger Koenker et al. “Adaptive L-estimation for linear models”. In : *The Annals of Statistics* 17.1 (1989), p. 362–381 (cf. p. 52).

[105] James L Powell. “Censored regression quantiles”. In : *Journal of econometrics* 32.1 (1986), p. 143–155 (cf. p. 45).

[106] Clémentine Prieur et Joanna Jongwane. “Mieux modéliser le climat grâce aux statistiques”. In : *Interstices* (nov. 2015) (cf. p. 14).

[107] Svetlozar T Rachev et Ludger Rüschendorf. *Mass Transportation Problems : Volume I : Theory*. T. 1. Springer Science & Business Media, 1998 (cf. p. 148).

[108] F. Riesz et B. Szoekefalvi-Nagy. *Lecons d'analyse fonctionnelle*. 1972 (cf. p. 35).

[109] Alix Rigal, Jean-Marc Azaïs et Aurélien Ribes. “Estimating daily climatological normals in a changing climate”. In : *Climate Dynamics* (2018), p. 1–12 (cf. p. 72, 135, 136, 140).

[110] Olivier ROUSTANT. “Mémoire d’Habilitation à Diriger des Recherches”. Thèse de doct. Citeseer, 2011 (cf. p. 27).

[111] Walter Rudin et al. *Analyse réelle et complexe : cours et exercices*. Dunod, 1998 (cf. p. 28).

[112] Saburou Saitoh et Yoshihiro Sawano. *Theory of reproducing kernels and applications*. Springer, 2016 (cf. p. 24).

[113] Simon C. Scherrer, Christof Appenzeller et Mark A. Liniger. “Temperature trends in Switzerland and Europe : Implications for climate normals”. In : *International Journal of Climatology* (2006) (cf. p. 75).

[114] Simon C Scherrer et al. “European temperature distribution changes in observations and climate change scenarios”. In : *Geophysical Research Letters* 32.19 (2005) (cf. p. 147).

[115] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. London : Chapman & Hall, 1986 (cf. p. 149).

[116] Stephen M Stigler. “Studies in the history of probability and statistics xl boscovich, simpson and a 1760 manuscript note on fitting a linear relation”. In : *Biometrika* 71.3 (1984), p. 615–620 (cf. p. 40).

[117] Thomas F Stocker et al. *Climate change 2013 : The physical science basis*. 2013 (cf. p. 17, 19).

[118] Karl E Taylor. “A summary of the CMIP5 experiment design”. In : *http ://cmip-pcmdi. llnl. gov/cmip5/docs/Taylor_CMIP5_design. pdf* (2009) (cf. p. 17).

[119] Karl E Taylor, Ronald J Stouffer et Gerald A Meehl. “An overview of CMIP5 and the experiment design”. In : *Bulletin of the American Meteorological Society* 93.4 (2012), p. 485–498 (cf. p. 18).

[120] C. Tebaldi et J. Arblaster. “Pattern scaling : Its strengths and limitations, and an update on the latest model simulations”. In : *Climatic Change* 122 (2014), p. 459–471 (cf. p. 81, 141).

[121] Claudia Tebaldi et al. “Pattern scaling : Its strengths and limitations, and an update on the latest model simulations”. In : *Climatic Change* 122.Issue 3 (2014), p. 459–471 (cf. p. 15).

[122] Detlef P Van Vuuren et al. “The representative concentration pathways : an overview”. In : *Climatic change* 109.1-2 (2011), p. 5 (cf. p. 18).

[123] R Vautard et al. “Extreme fall 2014 precipitation in the Cévennes mountains”. In : *Bulletin of the American Meteorological Society* 96.12 (2015), S56–S60 (cf. p. 133).

[124] Robert Vautard et al. “The European climate under a 2 C global warming”. In : *Environmental Research Letters* 9.3 (2014), p. 034006 (cf. p. 126).

[125] Konstantin Y Vinnikov et al. “Analysis of diurnal and seasonal cycles and trends in climatic records with arbitrary observation times”. In : *Geophysical Research Letters* 31.6 (2004), n/a–n/a (cf. p. 82, 141).

[126] A. Voldoire et al. “Evaluation of CMIP6 DECK experiments with CNRM-CM6-1”. In : *Journal of Advances in Modeling Earth Systems* 11 (juin 2019) (cf. p. 12).

[127] Grace Wahba. *Spline models for observational data*. T. 59. Siam, 1990 (cf. p. 30, 32).

[128] AM Walker. “A note on the asymptotic distribution of sample quantiles”. In : *Journal of the Royal Statistical Society : Series B (Methodological)* 30.3 (1968), p. 570–575 (cf. p. 43).

[129] D. S. Wilks. “Projecting "normals" in a nonstationary climate”. In : *Journal of Applied Meteorology and Climatology* 52.2 (2013), p. 289–302 (cf. p. 75, 78, 128, 169).

[130] Daniel S. Wilks et Robert E. Livezey. “Performance of Alternative “Normals” for Tracking Climate Changes, Using Homogenized and Nonhomogenized Seasonal U.S. Surface Temperatures”. In : *Journal of Applied Meteorology and Climatology* 52.8 (2013), p. 1677–1687. eprint : `https://doi.org/10.1175/JAMC-D-13-026.1` (cf. p. 75, 78, 79).

[131] Daniel S Wilks et Robert L Wilby. “The weather generation game : a review of stochastic weather models”. In : *Progress in physical geography* 23.3 (1999), p. 329–357 (cf. p. 133).

[132] DanielS Wilks. ""The stippling shows statistically significant grid points" : How research results are routinely overstated and overinterpreted, and what to do about it". In : *Bulletin of the American Meteorological Society* 97.12 (2016), p. 2263–2273 (cf. p. 127).

[133] DS Wilks. "Resampling hypothesis tests for autocorrelated fields". In : *Journal of Climate* 10.1 (1997), p. 65–82 (cf. p. 127).

[134] Christopher KI Williams et Carl Edward Rasmussen. *Gaussian processes for machine learning*. T. 2. 3. MIT press Cambridge, MA, 2006 (cf. p. 111).

[135] S. N. Wood. "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models". In : *Journal of the Royal Statistical Society (B)* 73.1 (2011), p. 3–36 (cf. p. 135).

[136] Simon N Wood, Natalya Pya et Benjamin Säfken. "Smoothing parameter and model selection for general smooth models". In : *Journal of the American Statistical Association* 111.516 (2016), p. 1548–1563 (cf. p. 170).

[137] S.N Wood. *Generalized Additive Models : An Introduction with R*. 2e éd. Chapman et Hall/CRC, 2017 (cf. p. 135).

[138] World Meteorological Organization. "The Role OF Climatological Normals In a Changing Climate". In : *WCDMP-No. 61* WMO-TD No.1377.World Climate Data and Monitoring Programme (2007) (cf. p. 75).

[139] William Wright. "Discussion paper on the calculation of the standard climate normals : A proposal for a dual system". In : *World Climate Data and Monitoring Program, accessed* 14 (2014) (cf. p. 75).

[140] Yuhong Yang. "Can the strengths of AIC and BIC be shared ? A conflict between model indentification and regression estimation". In : *Biometrika* 92.4 (déc. 2005), p. 937–950. eprint : `http://oup.prod.sis.lan/biomet/article-pdf/92/4/937/1099625/924937.pdf` (cf. p. 66).

[141] Thomas W. Yee. *Vector Generalized Linear and Additive Models : With an Implementation in R*. New York, USA : Springer, 2015 (cf. p. 135).

[142] Thomas W. Yee et C. J. Wild. “Vector Generalized Additive Models”. In : *Journal of Royal Statistical Society, Series B* 58.3 (1996), p. 481–493 (cf. p. 135).

[143] Jinlun Zhang, D Andrew Rothrock et Michael Steele. *Projections of an Ice-Diminished Arctic Ocean-Retrospection and Future Projection*. 2017 (cf. p. 16).

[144] X Zhang et al. “Changes in temperature and precipitation across Canada ; Chapter 4 in Bush E, Lemmen DS.(Eds.) Canada’s Changing Climate Report”. In : *Government of Canada, Ottawa, Ontario* (2019), p. 112–193 (cf. p. 103).

Résumé —

Les normales climatiques sont habituellement calculées comme des moyennes sur une période observée de 30 ans. Dans un contexte de changement climatique, ces normales, même ré-évaluées régulièrement, sont « en retard » sur le climat présent. Le premier objectif de ce travail de thèse est d'estimer des normales climatiques non-stationnaires, dans le but de disposer d'une référence non-biaisée pour le climat présent. Une bonne propriété pour de telles normales, considérées au pas de temps quotidien, est de présenter une certaine régularité, à la fois en terme de cycle annuel et vis-à-vis du changement climatique. Pour cette raison, l'estimation de ces normales sera basée sur des techniques de lissage spline telles que proposées dans Azaïs et Ribes (2016). La modélisation proposée, inspirée d'une hypothèse de "pattern scaling", permettra l'étude de la dérive saisonnière due au changement climatique.

Au-delà de la seule valeur moyenne (normale), le climat - considéré à un site et à une date donnés - se caractérise par une distribution de valeurs possibles. Un prolongement naturel de l'estimation de normales consiste à estimer l'ensemble de cette distribution, avec une contrainte de régularité sur la forme de celle-ci. Il s'agit alors de proposer une forme de régression quantile régularisée. On obtient ainsi, pour un paramètre donné, une description fine du climat en un site donné, et de son cycle annuel.

Ces deux aspects, nous amènent à ré-examiner la théorie des RKHS (Espace de Hilbert à noyau reproduisant) ainsi que celle de la régression quantile. La complexité des modèles considérés, autant dans le cas des normales que des distributions, fait l'objet d'un examen minutieux.
Enfin, nous proposons d'utiliser ces résultats pour revisiter et améliorer la description des changements climatiques passés et futurs, par exemple via l'utilisation d'analogues climatiques.

Mots clés : normales climatiques non-stationnaires, splines de lissage, RKHS, régression quantile, analogues climatiques, statistique des changements climatiques.

Abstract —

Climate normals are usually calculated as averages over a 30-year observational period. In the context of a changing climate, these normals, even re-evaluated frequently, are "lagging behind" the current climate. The primary objective of this thesis is the estimation of non-stationary normals, in order to acquire an unbiased reference of the present climate. A good property for such normals, which are considered at the daily timescale, is to possess a certain regularity, both on the seasonal component and the secular change. For this reason, normal estimation shall be based on smoothing spline techniques such as proposed by Azaïs et Ribes (2016). The proposed modeling, inspired by a "pattern scaling" assumption, enables the study of the seasonal drift due to climate change.

Beyond the mean value (normals), climate considered at a given date and location can be characterized by its distribution of possible values. A natural extension of normal estimation, would be to estimate the entire distribution with regularity constraints on its shape. We thus propose to address this question in a regularized quantile regression framework. In that respect and for a given climate variable, we obtain a detailed description of climate, at a given localization, and of its annual cycle.

Both of these aspects lead to a re-examination of the RKHS (Reproducing Kernel Hilbert Space) and quantile regression theories. The complexity of the considered models is equally considered for normals and distribution and is meticulously examined. Lastly, we propose to exploit the preceding results to revisit and improve the description of future and past changes in climate, for example via the use of climate analogues.

Keywords : daily climate normals, unbiased estimate, normals accounting for climate change, Smoothing splines

CNRM - GAME, UMR 3589
42, Av. Gaspard Coriolis
31057 Toulouse Cedex, FRANCE

Institut de Mathématiques de Toulouse, UMR 5219
Université Toulouse III - Paul Sabatier
118 route de Narbonne, 31062 TOULOUSE CEDEX 9