



Générateur stochastique de temps multisite basé sur un champ gaussien multivarié

Marc Bourotte

► To cite this version:

Marc Bourotte. Générateur stochastique de temps multisite basé sur un champ gaussien multivarié. Mathématiques générales [math.GM]. Université d'Avignon, 2016. Français. NNT : 2016AVIG0415 . tel-01393617

HAL Id: tel-01393617

<https://theses.hal.science/tel-01393617>

Submitted on 23 Nov 2016

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.



Région
Provence
Alpes
Côte d'Azur



École doctorale n° 536 : Agros sciences et Sciences

THÈSE

pour obtenir le grade de docteur délivré par

l'Université d'Avignon et des Pays de Vaucluse

Spécialité doctorale "Biostatistique"

présentée et soutenue publiquement par

Marc BOUROTTE

le 17 Juin 2016

Générateur stochastique de temps multisite basé sur un champ gaussien multivarié

Jury

M. Denis ALLARD	INRA	Directeur de thèse
Mme Liliane BEL	AgroParisTech	Co-directrice de thèse
M. Jean-Noël BACRO	Université de Montpellier II	Rapporteur
M. Étienne LEBLOIS	IRSTEA	Rapporteur
Mme Céline LACAUX	Université d'Avignon et des Pays de Vaucluse	Examinatrice
Mme Valérie MONBET	Université de Rennes 1	Examinatrice

Biostatistique et Processus Spatiaux (BioSP)

INRA - Domaine Saint-Paul

Site Agroparc 84914 Avignon Cedex 9, France

T
H
È
S
E

Résumé

Les générateurs stochastiques de temps sont des modèles numériques capables de générer des séquences de données climatiques de longueur souhaitée avec des propriétés statistiques similaires aux données observées. Ces modèles sont de plus en plus utilisés en sciences du climat, hydrologie, agronomie. Cependant, peu de générateurs permettent de simuler plusieurs variables, dont les précipitations, en différents sites d'une région. Dans cette thèse, nous proposons un modèle original de générateur stochastique basé sur un champ gaussien multivarié spatio-temporel.

Un premier travail méthodologique a été nécessaire pour développer un modèle de covariance croisée entièrement non séparable adapté à la nature spatio-temporelle multivariée des données étudiées. Cette covariance croisée est une généralisation au cas multivarié du modèle non séparable spatio-temporel de Gneiting dans le cas de la famille de Matérn. La démonstration de la validité du modèle et l'estimation de ses paramètres par maximum de vraisemblance par paires pondérées sont présentées. Une application sur des données climatiques démontre l'intérêt de ce nouveau modèle vis-à-vis des modèles existants.

Le champ gaussien multivarié permet la modélisation des résidus des variables climatiques (hors précipitation). Les résidus sont obtenus après normalisation des variables par des moyennes et écarts-types saisonniers, eux-mêmes modélisés par des fonctions sinusoïdales. L'intégration des précipitations dans le générateur stochastique nécessite la transformation d'une composante du champ gaussien par une fonction d'anamorphose. Cette fonction d'anamorphose permet de gérer à la fois l'occurrence et l'intensité des précipitations. La composante correspondante du champ gaussien correspond ainsi à un potentiel de pluie, corrélé aux autres variables par la fonction de covariance croisée développée dans cette thèse.

Notre générateur stochastique de temps a été testé sur un ensemble de 18 stations réparties en zone à climat méditerranéen (ou proche) en France. La simulation conditionnelle et non conditionnelle de variables climatiques journalières (températures minimales et maximales, vitesse moyenne du vent, rayonnement solaire et précipitation) pour ces 18 stations soulignent les bons résultats de notre modèle pour un certain nombre de statistiques.

Mots clés : Générateur stochastique de temps, variables climatiques, champ Gaussian multivarié, processus spatio-temporels, fonction de covariance croisée, séparabilité, covariance de Matérn, vraisemblance par paires.

Abstract

Stochastic weather generators are numerical models able to simulate sequences of weather data with similar statistical properties than observed data. However, few of them are able to simulate several variables (with precipitation) at different sites from one region. In this thesis, we propose an original model of stochastic generator based on a spatio-temporal multivariate Gaussian random field.

A first methodological work was needed to develop a completely non separable cross-covariance function suitable for the spatio-temporal multivariate nature of studied data. This cross-covariance function is a generalization to the multivariate case of spatio-temporal non-separable Gneiting covariance in the case of the family of Matérn. The proof of the validity of the model and the estimation of its parameters by weighted pairwise maximum likelihood are presented. An application on weather data shows the interest of this new model compared with existing models.

The multivariate Gaussian random field allows the modeling of weather variables residuals (excluding precipitation). Residuals are obtained after normalization of variables by seasonal means and standard deviations, themselves modeled by sinusoidal functions. The integration of precipitation in the stochastic generator requires the transformation of a component of the Gaussian random field by an anamorphosis function. This anamorphosis function can manage both the occurrence and intensity of precipitation. The corresponding component of the Gaussian random field corresponds to a rain potential, correlated with other variables by the cross-covariance function developed in this thesis.

Our stochastic weather generator was tested on a set of 18 stations distributed over the Mediterranean area (or close) in France. The conditional and non-conditional simulation of daily weather variables (maximum and minimum temperature, average wind speed, solar radiation and precipitation) for these 18 stations show good result for a number of statistics.

Keywords : Stochastic weather generator, weather variables, multivariate Gaussian random field, spatio-temporal processes, cross covariance function, separability, Matérn covariance, pairwise likelihood.

Table des matières

1	Introduction	2
1.1	Temps, climat et météo	3
1.2	Données météorologiques	5
1.3	Générateurs stochastiques de temps	10
1.4	Plan de la thèse	13
1.5	Références	15
2	Rappels sur les champs gaussiens	18
2.1	Champs gaussiens et modèles de covariance associés	19
2.2	Méthodes d'inférence par vraisemblance pour les champs gaussiens	30
2.3	Références	36
3	Fully Nonseparable Cross-Covariance Functions for Multivariate Spatio-Temporal Data	39
3.1	Introduction	40
3.2	A class of non-separable space-time cross-covariance functions	44
3.3	Estimation	47
3.4	Simulation study	49
3.5	Western France weather dataset	58
3.6	Discussion	64
3.7	Références	67
4	Générateur stochastique de temps multisite	71
4.1	Introduction	72
4.2	Présentation de la zone d'étude et des données	73
4.3	Le modèle	76
4.4	Algorithme de simulation	90
4.5	Validation du générateur stochastique de temps	93
4.6	Conclusion et perspectives	104
4.7	Références	107
A	Annexes	I

Liste des figures

1.1	Carte des climats de France métropolitaine hors Corse en 2010.	7
1.2	Exemple de données journalières de température.	8
2.1	Relations entre les propriétés des fonctions de covariance.	26
3.1	Exemple de covariance spatio-temporelle de Gneiting.	42
3.2	Champ gaussien spatio-temporel bivarié avec fonction de covariance croisée Gneiting-Matérn.	46
3.3	Boxplots des scores de prédiction pour les 4 différents modèles.	57
3.4	Localisation des 13 stations.	58
3.5	Variogrammes spatial et temporel de la température.	59
3.6	Adéquation de la fonction de covariance croisée avec u fixé.	61
3.7	Adéquation de la fonction de covariance croisée avec h fixé.	62
3.8	Prédiction par simulation conditionnelle du rayonnement solaire, de la tem- pérature et de l'humidité.	63
4.1	Zone d'étude.	74
4.2	Densité du rayonnement solaire.	76
4.3	Modélisation de la saisonnalité par des fonctions sinusoïdales.	79
4.4	Principe du processus de pluie à l'aide d'un champ gaussien latent.	81
4.5	QQ-plots des précipitations en janvier.	84
4.6	Adéquation entre les covariances empiriques des résidus et le modèle. (1/2)	89
4.7	Adéquation entre les covariances empiriques des résidus et le modèle. (2/2)	90
4.8	Représentation de la densité des températures minimales aux 18 sites. . . .	94
4.9	Représentation de la densité des températures maximales aux 18 sites. . . .	94
4.10	Représentation de la densité des vitesses moyennes de vent aux 18 sites. . .	95
4.11	Représentation de la densité du rayonnement solaire aux 18 sites.	95
4.12	Nombre d'événements pluvieux en fonction de leur durée aux 18 sites. . . .	97
4.13	Nombre d'événements secs en fonction de leur durée aux 18 sites.	97
4.14	Nombre de stations dépassant simultanément le seuil de précipitation 8 mm.	98
4.15	Corrélations entre chaque paire de variables pour les 18 stations	99

4.16 Simulation conditionnelle des températures minimales et maximales, de la vitesse du vent et du rayonnement solaire à Avignon.	102
4.17 Simulation conditionnelle de l'occurrence et de l'intensité des précipita- tions à Avignon.	103
A.1 QQ-plots des précipitations en juillet.	XIII

Liste des tableaux

2.1	Fonctions complètement monotones.	30
2.2	Fonctions positives avec une dérivée complètement monotone.	30
3.1	Moyenne de la somme des variances estimées de tous les paramètres pour différents $\mathbf{d} = (d_S, d_T)$	50
3.2	Estimation de la moyenne et de l'écart type du paramètre de séparabilité pour différents $\mathbf{d} = (d_S, d_T)$	51
3.3	Résumé statistique des estimations des paramètres.	52
3.4	Comparaison de la performance d'estimation des différentes vraisemblances.	53
3.5	Scores de prédiction RMSE, MAE, CRPS et LogS des différentes vraisem- blances.	56
3.6	Résultats de la procédure d'estimation.	60
3.7	Comparaison des scores de prédiction.	61
3.8	Coefficients de covariance croisée empiriques et estimés.	61
4.1	Proportion de jours secs pour chacune des stations.	75
4.2	Estimation des paramètres de la fonction Ψ	82
4.3	Résultats de la procédure d'estimation des paramètres de la fonction de co- variance croisée.	87
4.4	Coefficients de covariance croisée empiriques et estimés.	87
4.5	Statistiques sur les cumuls mensuels agrégés des précipitations de la zone d'étude.	97
A.1	Estimation des paramètres de la fonction Ψ	XIV

Chapitre 1

Introduction

Sommaire

1.1 Temps, climat et météo	3
1.2 Données météorologiques	5
1.2.1 Particularités des données météorologiques	5
1.2.2 Présentation des jeux de données retenus	8
1.3 Générateurs stochastiques de temps	10
1.3.1 Propriétés	10
1.3.2 Intérêts	11
1.3.3 Modèles historiques	12
1.4 Plan de la thèse	13
1.5 Références	15

L'observation et la prévision du temps sont une préoccupation ancestrale pour l'être humain. Depuis des millénaires, les hommes ont intégré le concept de saisons et leurs alternances, dues à l'inclinaison de l'axe des pôles combinée à la révolution de la Terre autour du Soleil. D'ailleurs le terme climat provient du grec *klima* qui désigne l'inclinaison de la Terre par rapport au Soleil.

A la suite de la dernière glaciation, le début de la période néolithique voit naître l'agriculture au Moyen-Orient (10 000 av. J.-C.). Le radoucissement climatique favorise la croissance des plantes notamment des céréales sauvages abondantes dans cette partie du monde. Les premières traces écrites de relevés de données climatiques (et même de prévisions) datent de la première moitié du 1^{er} millénaire av. J.-C., et il y a fort à parier que dès le début de l'agriculture les hommes ont observé, relevé les phénomènes climatiques dans le but d'établir des stratégies adaptées à ces aléas. Même s'ils sont moins connus de nos jours, un grand nombre de proverbes populaires mêlant dates calendaires, conditions météorologiques et pratiques agricoles témoignent de ce lien fort. Dans la partie suivante, nous définirons plus précisément les notions de climat, temps et météorologie mais dans le langage courant, le temps, les conditions climatiques ou météorologiques expriment bien souvent la même idée.

A partir du 17^{ième} siècle et plus particulièrement au cours du 19^{ième} siècle, les premiers réseaux météorologiques apparaissent en Europe et aux Etats-Unis avec le développement des instruments de mesure comme l'anémomètre (XV^e) ou le thermomètre (XVII^e). A la même époque (XVIII^e) les statistiques apparaissent notamment avec la construction des premières tables de mortalité. Cette branche des mathématiques possède une composante théorique s'appuyant sur les probabilités et une composante appliquée utilisant des données. La quantité de données enregistrées par les différentes stations météorologiques étant de plus en plus importante, il était naturel que les statisticiens s'intéressent à ces données.

Cette thèse s'inscrit dans ce lien fort entre agriculture, statistiques et données climatiques. Nous proposons un modèle statistique innovant permettant de générer facilement des données météorologiques journalières régionalisées nécessaires aux modèles agronomiques. La thèse a été financée par la région PACA (Provence-Alpes-Côte d'Azur) et par l'INRA (Institut National de la Recherche Agronomique) au travers du métaprogramme ACCAF (Adaptation au Changement Climatique de l'Agriculture et de la Forêt).

1.1 Temps, climat et météo

Cette partie a pour objectif de définir un peu plus précisément les différents termes tels que temps, climat ou météo qui peuvent être parfois confondus dans le langage courant.

Dans le glossaire du quatrième rapport d'évaluation (AR4, 2007) de l'IPCC (Intergo-

vernmental Panel on Climate Change) ou GIEC (Groupe d'experts Intergouvernemental sur l'Evolution du Climat) en français, le climat est défini de la manière suivante : "au sens étroit du terme, le climat désigne généralement le temps moyen ; il s'agit plus précisément d'une description statistique en fonction de la moyenne et de la variabilité de grandeurs pertinentes sur des périodes variant de quelques mois à des milliers, voire à des millions d'années (la période classique, définie par l'Organisation Météorologique Mondiale, est de 30 ans). Ces grandeurs sont le plus souvent des variables de surface telles que la température, les précipitations et le vent". Sur ces périodes, le climat peut être considéré comme un processus stochastique de plusieurs variables ou événements climatiques et donc être défini par des lois de probabilité - densités ou fonctions de répartition - d'événements climatiques. Ces événements sont caractérisés par leur occurrence et leur intensité ; l'évolution des températures, la vitesse et la direction du vent, les précipitations, des événements extrêmes comme les canicules, les tempêtes sont autant d'exemples d'événements climatiques. L'utilisation de lois de probabilités permet de prendre en compte la fluctuation du climat, ou variabilité naturelle, due aux échanges d'énergie entre l'atmosphère et les océans, à l'activité solaire ou encore à l'activité volcanique. Ainsi, le terme de temps (*weather* en anglais) prend toute sa signification et définit, au sens probabiliste, une réalisation (ou une trajectoire) du climat. Le climat dépend aussi de la zone géographique considérée ce qui veut dire que les probabilités associées à des événements climatiques peuvent varier spatialement. Le climat à l'échelle du globe est différent de celui d'un pays, qui lui même présente des hétérogénéités en fonction des régions qui composent ce pays.

Wikipédia fournit comme définition pour la météorologie : "science qui a pour objet l'étude des phénomènes atmosphériques tels que les nuages, les précipitations ou le vent dans le but de comprendre comment ils se forment et évoluent en fonction des paramètres mesurés tels que les pressions, la température et l'humidité". A première vue, cette définition ne diffère pas énormément de celle de la climatologie qui est l'étude du climat.

La notion d'échelle temporelle est le principal aspect qui distingue la météorologie de la climatologie. En effet, la météorologie se concentre sur l'étude et la prévision des mêmes variables (pressions, températures, précipitations, etc) sur une période relativement courte allant de quelques heures à quelques jours voire une à deux semaines maximum. De plus, les prévisions météorologiques s'intéressent fortement à la chronologie des événements. Si une prévision indique cinq millimètres de précipitation le matin et que cet événement se produit dans l'après midi, la prévision météorologique sera fausse. Alors que d'un point de vue statistique à l'échelle journalière, la distribution de l'occurrence et de l'intensité de précipitations sera la même dans les deux cas. Les prévisions météorologiques et climatiques sont obtenus à l'aide de modèles numériques qui s'appuient sur les équations de Navier-Stokes, la loi des gaz parfaits ou encore les équations de bilan de l'énergie pour modéliser l'atmosphère. Les prévisions climatiques nécessitent

aussi la prise en compte des océans, des banquises et de leur interactions avec l'atmosphère. Les projections de températures globales moyennes en 2100 réalisées en prenant en compte les scénarios émis par le GIEC sur les émissions de gaz à effet de serre sont des exemples de prévisions climatiques et fournissent des intervalles de température avec une probabilité associée.

Les Modèles de Circulation Générale (MCG, ou GCM en anglais) appartiennent à la famille des modèles climatiques et résolvent des équations primitives de la mécanique et de la thermodynamique des fluides géophysiques. Ils fournissent des simulations sous forme de grilles avec une résolution spatiale de la maille d'environ 250 km x 250 km. Les modèles climatiques régionaux (RCM) sont une version des GCMs travaillant à plus petite échelle puisqu'ils cherchent à résoudre explicitement les équations de la dynamique atmosphérique régionale. Dans ce cas, le maillage est plus fin : de 50 à environ 5 km. Cependant, au sein d'une même maille, différentes localisations, qui peuvent être matérialisées par des stations météorologiques, sont susceptibles de présenter des distributions d'événements climatiques légèrement différentes. Cette nuance ne peut être restituée fidèlement par les GCMs ou RCMs.

Le GCM français ARPEGE (Action de Recherche Petite Echelle Grande Echelle) développé par Météo-France permet la prévision jusqu'à quatre jours d'échéance sur les zones d'intérêt pour la France mais permet aussi dans une version "climat" de simuler les évolutions du climat aux longues échéances à fine échelle.

Pour terminer avec les définitions, dans cette thèse nous emploierons indifféremment les termes variables climatiques et météorologiques, et le terme temps (*weather*) fera référence à une réalisation de ces variables pour le climat de la zone considérée. Bien sûr le mot temps (*time*) sera aussi utilisé pour décrire la chronologie d'événements (la flèche du temps).

1.2 Données météorologiques

1.2.1 Particularités des données météorologiques

Les données jouent un rôle fondamental en statistique. Avant même d'envisager tout travail de modélisation, il est important d'analyser la nature des données et d'effectuer une étude de statistique descriptive pour décider de la direction à suivre. Des outils statistiques ne faisant pas intervenir d'hypothèse de distribution ou de modèle comme les boxplots, les fonctions d'autocorrélation ou les variogrammes sont très utiles et seront utilisés dans les chapitres suivants.

Les données météorologiques sont complexes à plusieurs niveaux. La quantité de données disponibles est importante ce qui est souvent un atout pour le statisticien mais complexifie aussi les méthodes qui seront utilisées pour les traiter. En effet, lorsque l'on étudie le climat d'une zone géographique, on considère un réseau de stations fournissant

les données de plusieurs variables météorologiques au cours d'une période d'une dizaine d'années au minimum. A titre d'exemple, les stations météorologiques professionnelles du réseau de Météo-France, appelé réseau Radome, sont au nombre de 554 en métropole (une tous les 30 km). On peut citer aussi EDF (300 stations), la direction générale de la prévention des risques (500 stations) et l'INRA (40 stations) qui échangent des données dans le cadre des réseaux d'intérêt commun (RIC).

La nature des données est un niveau supplémentaire de complexité. Naturellement, on entend par données les valeurs relevées par des appareils de mesure mais il existe aussi des données issues des modèles numériques comme les GCMs, RCMs présentés dans la section précédente, les générateurs stochastiques de temps (SWG) qui seront développés dans la section suivante ou encore les modules d'analyse comme SAFRAN [Quintana-Seguí et al., 2008] fournissant des données de réanalyse. De plus, un certain nombre de séries de données historiques, i.e. enregistrées, présentent des données manquantes. Ces "trous" peuvent être plus ou moins longs en fonction de la cause (panne d'un appareil de mesure, maintenance, intempéries ...) et peuvent être comblés par différentes méthodes quand c'est nécessaire. Par exemple, lorsqu'il manque ponctuellement une valeur, celle-ci peut être remplacée par une valeur moyenne (mensuelle, annuelle) ou par la valeur qui survient le plus souvent. Lorsque de plus longues séquences sont manquantes, des méthodes plus sophistiquées (régressions, classifications, Kotsiantis et al. [2006]) essayent de recréer les dépendances temporelles voire des dépendances plus complexes comme celles entre variables. Comme nous le verrons dans la partie suivante, certains générateurs stochastiques de temps peuvent aussi remplir cette fonction, et ce sera le cas du modèle que nous présenterons dans cette thèse.

Enfin, la non homogénéité des données disponibles participent grandement à cette complexité. La non homogénéité peut provenir de différences d'échelles spatiale, temporelle ou de la nature des sources des données historiques.

Premièrement, les sorties des GCMs sont des grilles qui ont pour résolution spatiale un maillage de 250 km $\times$ 250 km (de 50 km à environ 5 km pour les RCMs) tandis que les stations fournissent des données ponctuelles. Ainsi, utiliser des données issues de différentes résolutions spatiales nécessitent un travail supplémentaire. C'est ce qui est proposé par exemple dans Bechler et al. [2015] pour la simulation conditionnelle de précipitations extrêmes dans le Sud-Est de la France avec des données basse résolution (sorties de RCM, grille de résolution 50 km et 15 mailles couvrant la zone d'intérêt) et haute résolution (données SAFRAN, 457 points répartis uniformément). De même, le pas de temps est variable selon les stations et les variables, allant de quelques minutes à journalier.

Deuxièmement, l'hétérogénéité des données météorologiques provient aussi des sources puisque les différents acteurs des réseaux météorologiques n'utilisent pas forcément les mêmes méthodes de récolte des données. Par exemple, les stations INRA mesurent la vitesse du vent à une hauteur de 2 mètres tandis que les stations Météo-France mesurent à

La carte de France est divisée en zones climatiques colorées, avec des villes principales indiquées par des carrés noirs. Des flèches indiquent les influences des vents d'ouest et du Mistral.

Clés de lecture :

- Climat océanique nord-ouest
- Climat océanique aquitain
- Climat océanique dégradé
- Climat semi-continental
- Climat méditerranéen (Köppen Csa-Csb)
- Climat de montagne ou sous influence montagnarde

Annotations sur la carte :

- VENTS D'OUEST (flèche pointant vers le nord-ouest)
- VENTS D'OUEST (flèche pointant vers le sud-ouest)
- MISTRAL (flèche pointant vers le sud-est)
- TRAMONTANE (flèche pointant vers le sud-est)

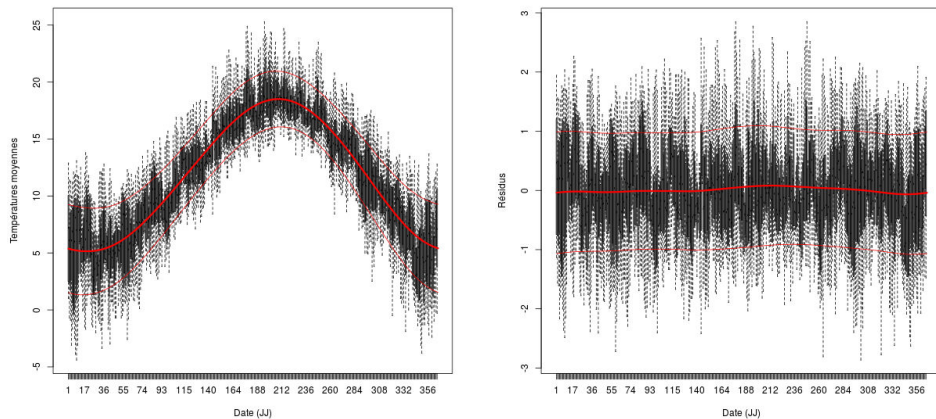


FIGURE 1.2 – A gauche : courbe des températures moyennes journalières au Rheu (Bretagne) de 2003 à 2012. Chaque jour calendaire est représenté par un boxplot. La courbe rouge est un lissage par spline. A droite : résidus des mêmes températures, i.e. après normalisation.

1.2.2 Présentation des jeux de données retenus

Dans cette thèse nous travaillerons avec des données fournies par le portail Climatik (géré par l'unité Agroclim) qui est l'outil de mise à disposition des données agroclimatiques de la base de l'INRA aux utilisateurs. Cette base contient les données des 40 stations INRA ainsi qu'un nombre équivalent de stations Météo-France. Plusieurs variables météorologiques et agroclimatiques sont disponibles à différents pas de temps (suivant les variables). Nous avons sélectionné les variables journalières suivantes :

- les températures (minimales, maximales et moyennes) exprimées en degré Celsius (°C) et mesurées par un thermomètre ;
- le rayonnement solaire global exprimé en Joule par centimètre carré (J/cm^2) et mesuré par un pyranomètre ;
- l'humidité moyenne exprimée en pourcentage (%) et mesurée par un hygromètre ;
- les précipitations exprimées en millimètre (mm) et mesurées par un pluviomètre ;
- la vitesse moyenne du vent exprimée en mètre par seconde et mesurée par un anémomètre (m/s).

Plusieurs remarques s'imposent concernant ces diverses variables météorologiques. Premièrement, du fait de la précision des instruments, toutes ces variables sont discrètes. Par exemple, la précision est de l'ordre de l'unité pour le rayonnement et l'humidité tandis qu'elle est du dixième d'unité pour les températures et la vitesse du vent. Pour les précipitations, les jauges pluviométriques ont une précision de 0.5 mm. Deuxièmement, certaines variables sont bornées. C'est le cas du rayonnement solaire et de la vitesse du vent qui ne peuvent pas être négatifs ou de l'humidité qui est comprise entre 0 et 100. Les

précipitations sont un autre cas de variable bornée, ici positive, et qui présentent la particularité supplémentaire de comporter une proportion très importante de zéros (*zero-inflated* en anglais) correspondant à l'absence de pluie. Les précipitations constituent une variable spéciale en statistique du climat et elles seront détaillées plus précisément dans le chapitre 4.

Dans le chapitre 3, seules les trois variables journalières rayonnement solaire, température moyenne et humidité sont utilisées. Nous n'avons travaillé qu'avec des stations INRA car l'humidité n'était pas disponible pour les stations Météo-France de la base Climatik. Toutes les stations retenues pour ce travail sont réparties sur la façade ouest (plus ou moins dans les terres) de la France allant de la côte basque à la pointe nord de la Bretagne. Comme nous pouvons le voir sur la figure 1.1, cette zone géographique présente une bonne homogénéité climatique et correspond à un climat océanique. Deux des caractéristiques de la zone sont la faiblesse de l'écart de température moyenne entre l'hiver et l'été, et la dominance des vents d'ouest. La période d'étude s'étend de 2002 à 2013 et nous nous sommes concentrés sur les mois de janvier pour éviter les effets saisonniers. Ceci est une stratégie classique régulièrement utilisée comme dans [Bessac et al. \[2015\]](#) avec des données de vitesse de vent au nord-ouest de la Bretagne pour la conception d'un générateur stochastique multi-sites de vent. Le graphique de gauche de la figure 1.2 représente les températures moyennes journalières (bâtonnets noirs) de 2002 à 2013 au Rheu à côté de Rennes (Bretagne). La courbe rouge correspond à une moyenne lissée par splines et nous pouvons remarquer que la courbe pour les 31 premiers jours calendaires, i.e. le mois de janvier, est relativement peu incurvée ce qui indique une faible influence saisonnière sur cette période. Deux hypothèses sont alors faites : l'influence saisonnière est négligée pendant le mois de janvier et les différents mois de janvier de la série sont considérés indépendants. Cette dernière hypothèse est importante puisque ces différents mois de janvier seront considérés comme les réalisations d'un même processus stationnaire (notion qui sera plus explicitement abordée dans la section 2.1).

Dans le chapitre 4, nous présenterons l'application de notre générateur stochastique de temps sur des données issues de stations à climat méditerranéen. La classification de Köppen, utilisée pour la carte 1.1, est une classification des climats fondée sur les précipitations et les températures. Le climat méditerranéen est un climat tempéré caractérisé par une saison sèche en été. D'autres zones géographiques dans le monde se situent en climat méditerranéen comme la Californie, le centre du Chili, la région du Cap en Afrique du sud ou le sud ouest de l'Australie. Les 18 stations utilisées dans cette étude sont toutes situées à l'ouest de Marseille et sont exposées aux vents de nord comme le Mistral et la Tramontane. Les variables météorologiques journalières considérées sont au nombre de cinq : le rayonnement solaire, les températures minimales et maximales, la vitesse moyenne du vent et les précipitations. Initialement, nous voulions également prendre en compte l'humidité relative comme dans le chapitre 3, mais il aurait fallu travailler uniquement avec

les stations INRA de la zone d'intérêt (11 stations). Or l'objectif était de travailler avec un nombre plus important de variables et de stations que dans le chapitre 3, et donc nous avons considéré les stations INRA et Météo-France disponibles dans notre zone d'intérêt - région de France à climat méditerranéen.

1.3 Générateurs stochastiques de temps

1.3.1 Propriétés

Les générateurs stochastiques de temps (*Stochastic Weather Generator* (SWG) en anglais) appartiennent à la famille des modèles numériques capables de générer des données météorologiques comme les GCMs ou les RCMs. Mais contrairement à ces derniers qui s'appuient sur une approche déterministe des lois physico-chimiques de l'atmosphère, les SWGs sont basés sur une approche statistique et probabiliste du climat. Plus spécifiquement, un générateur stochastique de temps est un modèle statistique qui permet de simuler des séquences de données de longueur souhaitée avec des propriétés statistiques similaires aux données observées. Il résulte de cette définition qu'il existe de nombreux modèles suivant les critères définis par le modélisateur : monosites ou multisites, univariés ou multivariés. De même, les propriétés statistiques visées peuvent différer selon les modèles.

La majorité des générateurs sont conçus pour reproduire fidèlement les propriétés d'espérance et de variance des distributions empiriques des données historiques. Mais au delà de ces caractéristiques, la corrélation temporelle reçoit toujours une grande attention puisque les données sont des séries temporelles. Un certain nombre d'autres corrélations peuvent également être étudiées en fonction de la nature des observations : la corrélation spatiale pour des données multisites ou encore la corrélation entre variables pour des données multivariées. Ces trois types de corrélation peuvent être qualifiées de directes et il existe aussi des corrélations croisées entre elles, qui peuvent être prises en compte par le modélisateur. Toutes ces considérations permettent de reproduire le plus scrupuleusement possible la dynamique spatio-temporelle des données historiques.

Un autre champ de la statistique s'intéresse aux propriétés des valeurs appartenant aux queues des distributions et se nomme théorie des valeurs extrêmes [De Haan and Ferreira, 2007]. On comprend rapidement l'intérêt de générer fidèlement ces valeurs extrêmes comme pour le dimensionnement d'ouvrages d'art (ponts, barrages, etc) dans le cas d'inondations. Ainsi deux approches peuvent être envisagées : la première consiste en l'analyse des maxima par intervalles de temps fixes (annuels, décennales) à l'aide de la loi généralisée des extrêmes ou GEV (*Generalized Extreme Value*) ; la seconde se base sur l'analyse des valeurs au-dessus d'un seuil selon la méthode des dépassements ou POT (*Peak Over Threshold*) et utilise la loi généralisée de Pareto ou GPD (*Generalized Pareto Distribution*). Les modèles avancés essayent de mixer les apports des diverses théories

pour représenter au mieux le corps de la distribution d'une variable ainsi que sa queue. C'est ce qui est proposé dans [Vrac and Naveau \[2007\]](#) avec un générateur stochastique de précipitation permettant de simuler des pluies fines tout comme des orages. Le modèle est basé sur le mélange d'une loi Gamma et d'une loi de Pareto généralisée. En effet, la loi Gamma est réputée pour approximer correctement les précipitations [\[Bellone et al., 2000\]](#) à l'exception des valeurs extrêmes [\[Katz et al., 2002\]](#) du fait de sa queue de distribution trop légère. Ce manque est corrigé par l'apport de la GPD.

Enfin, dans certains cas ce n'est plus la valeur d'une variable à proprement parler qui est le plus important mais la durée pendant laquelle ses valeurs restent au dessus d'un certain seuil. Les vagues de froid ou de chaleur (canicules) sont des exemples météorologiques de ce type de phénomènes. Même si la définition en terme de températures diffère régionalement, les canicules sont caractérisées par une baisse significative de l'amplitude thermique entre le jour et la nuit. Les sécheresses sont encore un exemple et sont caractérisées par la durée de non-occurrence de précipitation. Ici du point de vue mathématique, la durée est discrète et peut être modélisée par des distributions de type Poisson ou binomiale négative [\[Rydén, 2014\]](#). [Semenov \[2007\]](#) démontre les capacités du générateur LARS-WG [\[Semenov et al., 1998\]](#) à reproduire correctement la durée des événements de vague de chaleur pour 20 sites avec divers climats. Il sera intéressant de voir si le modèle que nous proposons est capable de reproduire de tels événements qui ont de forts impacts sur l'agriculture et les activités humaines en général.

1.3.2 Intérêts

Même si le concept de générateur stochastique n'est pas réservé aux séries météorologiques, il présente de nombreux intérêt dans ce domaine, notamment vis-à-vis des modèles physiques. Les SWGs permettent de simuler rapidement des séries de longueurs choisies (même très longues) grâce à des algorithmes peu coûteux en terme computationnel puisqu'ils peuvent tourner sur des ordinateurs personnels. Ces séries simulées peuvent être utilisées à plusieurs effets.

L'utilisation la plus fréquente des séries simulées par un SWG est de servir de variables d'entrée d'un autre modèle (hydrologique, agronomique ou économique par exemple). En effet les données historiques disponibles ne sont pas toujours assez nombreuses ou enregistrées sur une durée assez longue pour faire fonctionner ou tester correctement ces modèles. C'est notamment le cas lorsque l'on veut effectuer une analyse de sensibilité (méthode permettant d'analyser un modèle mathématique en étudiant l'impact de la variabilité des facteurs d'entrée du modèle sur la variable de sortie) ou une analyse d'impact, qui requièrent une quantité importante de données. La thèse de Cédric Flecher (chapitre 4, [Flecher \[2009\]](#)) fournit un exemple d'utilisation couplée d'un SWG, ici le WACS-Gen (pour Weather-state Approach Conditionnaly Skew-normal Generator [\[Flecher et al., 2010\]](#) modèle monosite permettant de générer températures, rayonne-

ment, précipitations et vent), avec un modèle agronomique STICS [Brisson et al., 2003] de culture de blé et de maïs.

Parfois, certains SWGs peuvent répondre aux problèmes de données manquant partiellement dans une série ou de données non disponibles pour un site. Le générateur proposé au chapitre 4 appartient à cette classe de modèles.

Enfin plus récemment, le contexte de changement climatique offre une portée supplémentaire aux SWGs. Plutôt que d'opposer SWGs et GCMs/RCMs, plusieurs travaux [Semenov and Barrow, 1997][Vrac and Naveau, 2007][Wilks, 2010][Wilks, 2012][Bechler et al., 2015] montrent la complémentarité de ces modèles, qui trouve tout son intérêt dans un contexte de *downscaling*. En effet les GCMs/RCMs permettent de simuler des scénarios d'évolution du climat à l'échelle du globe ou de régions mais n'offrent pas d'informations climatiques à beaucoup plus fine échelle nécessaires aux analyses d'impact. Cette information à large échelle est transférée aux paramètres du générateur stochastique qui peut ainsi générer localement des données prenant en compte l'évolution du climat.

1.3.3 Modèles historiques

Dans cette section, nous présenterons rapidement quelques modèles "historiques" de générateurs stochastiques de temps. Pour une présentation plus exhaustive des modèles existants, nous renvoyons le lecteur vers les deux articles de review suivants : Wilks and Wilby [1999] et Ailliot et al. [2015]. Une approche naturelle consiste à tirer de manière "intelligente" les nouvelles données dans un ensemble de données historiques. Cette méthode dite par ré-échantillonnage a l'avantage de ne pas nécessiter d'hypothèses fortes comme celle de travailler dans un cadre gaussien, et les relations entre variables sont mieux respectées. Le SWG présenté dans Rajagopalan and Lall [1999], utilisant les k plus proches voisins (k -NN) et des distributions non paramétriques, est un exemple de générateur multivarié monosite par ré-échantillonnage. Yiou [2014] propose la prise en compte d'analogues de circulation atmosphérique (pression au niveau de la mer) à l'échelle de l'Europe pour générer spatialement les températures et les vagues de froid et chaleur. Une autre technique de ré-échantillonnage, appelée échantillonnage direct (*direct sampling*), permet de mieux prendre en compte les dépendances temporelles et est utilisée dans Oriani et al. [2014] pour simuler des précipitations journalières en Australie. Cependant l'approche par ré-échantillonnage présente le gros désavantage de ne pas simuler de nouvelles séquences quelque soit la technique utilisée (k -NN ou *direct sampling*).

L'approche paramétrique ou semi-paramétrique permet de simuler des séquences observées ou non et nécessite la construction de modèles probabilistes. Depuis le début, la pluie a été la variable qui a reçu le plus d'attention et les travaux de Gabriel and Neumann [1962] sur la modélisation d'occurrence de précipitations à Tel Aviv par une chaîne de Markov à deux états sont les prémices du développement des SWGs. Le modèle présenté dans Katz [1977] est l'un des premiers générateurs stochastiques de pluie.

Il s'appuie encore sur une chaîne de Markov d'ordre 1 à deux états (pluie et sec) et conditionnellement à ces états des montants de précipitations sont simulés indépendamment à l'aide d'une loi Gamma. Ce modèle monosite est aussi appelé *chain depend model*. Un peu plus tardivement, une version multisite est proposée par Wilks [1998]. L'aspect spatial de l'occurrence et du montant des précipitations est guidé par des variables normales (i.e gaussiennes) transformées. Ces travaux conduiront à un nouveau générateur stochastique de temps [Wilks, 1999] où les variables d'intérêt sont la pluie, les températures et le rayonnement. Un autre modèle (monosite) de référence, connu sous le nom de WGEN ou modèle de Richardson [Richardson, 1981] [Richardson and Wright, 1984], utilise aussi la dichotomie entre jours secs et pluvieux et simule conditionnellement à ces états de pluie les autres variables à l'aide de processus autorégressifs stationnaires gaussiens d'ordre 1 (AR(1)). Richardson [1981] présente aussi une méthodologie intéressante qui consiste à traiter les données en les normalisant à l'aide des moyennes et écarts-types saisonniers lissés (exemple de traitement : figure 1.2) pour travailler ensuite sur les résidus. Une partie de cette méthodologie sera reprise par un grand nombre de SWGs dont le nôtre. L'utilisation d'un processus autorégressif multivarié est reprise dans Kleiber et al. [2013] pour un générateur spatialisé qui est étendu à un générateur intégrant en plus les occurrences de précipitation via un modèle probit [Verdin et al., 2015]. La modélisation est plus sophistiquée puisqu'ils décomposent chaque variable (hors pluie) en une somme de deux éléments : une composante "climat local" modélisée par un processus autorégressif et d'une composante "temps" modélisée par un champ gaussien.

Comme nous avons pu le voir au travers de ce rapide tour d'horizon des modèles disponibles, les générateurs stochastiques de temps ont évolué de simples générateurs de pluie monosite à des générateurs spatialisés, multivariés ou parfois pouvant être centrés sur une variable d'intérêt précise (précipitations, vent). Les SWGs spatialisés pouvant simuler plusieurs variables dont les précipitations sont peu nombreux et sont souvent limités en nombre de variables ou par l'impossibilité de générer des données en des points où aucune donnée n'est disponible. Enfin il est difficile de s'affranchir des processus gaussiens lorsque l'on travaille avec plusieurs variables spatialisées.

1.4 Plan de la thèse

Cette introduction met en lumière la nécessité de développer des générateurs stochastiques de temps capables de simuler plusieurs variables météorologiques, dont les précipitations, en plusieurs sites d'une même région. Nous avons pris dès le départ la décision de s'appuyer sur les champs gaussiens pour construire notre modèle de SWG.

Le chapitre 2 présente les différentes propriétés des champs gaussiens et des fonctions

de covariance associées. Deux modèles de covariance (Matérn et Gneiting) sont également détaillés. Enfin ce chapitre se termine sur les approches d'estimation par vraisemblance.

Le chapitre 3 est consacré au développement méthodologique de modèles de covariance adaptés aux données spatio-temporelles multivariées pour des champs gaussiens stationnaires. Cette famille de fonctions de covariance peut être vue comme l'extension au cas multivarié des fonctions de covariance spatio-temporelles non séparables de Gneiting dans un cadre particulier :

- la fonction ψ , positive avec ses dérivées complètement monotones, ne dépend pas des variables ;
- la fonction φ complètement monotone est soit une covariance de Matérn ou soit une covariance de Cauchy ;
- la fonction φ dépend des variables au travers de sa paramétrisation comme pour les modèles de Matérn multivariés.

Une démonstration de la validité de tels modèles sera apportée ainsi que la capacité de l'inférence par maximum de vraisemblance par paires à estimer correctement l'ensemble des paramètres. Une étude de simulations ainsi qu'une application sur des données météorologiques compléteront ce chapitre.

Le chapitre 4 décrit la construction du SWG étape par étape. La modélisation des précipitations par un champ gaussien latent est développée ainsi que les modifications résultantes des équations de vraisemblance. Enfin la validité du générateur est testée sur un jeu de données météorologiques provenant de stations réparties en zone à climat méditerranéen.

1.5 Références

- Ailliot, P., Allard, D., Monbet, V., and Naveau, P. (2015). Stochastic weather generators : an overview of weather type models. *Journal de la Société Française de Statistique*, 156(1) :101–113. [12](#)
- Bechler, A., Vrac, M., and Bel, L. (2015). A spatial hybrid approach for downscaling of extreme precipitation fields. *Journal of Geophysical Research : Atmospheres*. [6](#), [12](#)
- Bellone, E., Hughes, J. P., and Guttorp, P. (2000). A hidden markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Climate research*, 15(1) :1–12. [11](#)
- Bessac, J., Ailliot, P., and Monbet, V. (2015). Gaussian linear state-space model for wind fields in the north-east atlantic. *Environmetrics*, 26(1) :29–38. [9](#)
- Brisson, N., Gary, C., Justes, E., Roche, R., Mary, B., Ripoche, D., Zimmer, D., Sierra, J., Bertuzzi, P., Burger, P., et al. (2003). An overview of the crop model stics. *European Journal of agronomy*, 18(3) :309–332. [12](#)
- De Haan, L. and Ferreira, A. (2007). *Extreme value theory : an introduction*. Springer Science & Business Media. [10](#)
- Flecher, C. (2009). *Conception d'un générateur de temps : développements méthodologiques pour des distributions asymétriques multivariées; application à l'étude de sensibilité des cultures à la variabilité du climat et à ses modifications*. PhD thesis, <http://informatique-mia.inra.fr/biosp/sites/ciam.inra.fr.biosp/files/TheseCFlecher.pdf>. [11](#)
- Flecher, C., Naveau, P., Allard, D., and Brisson, N. (2010). A stochastic daily weather generator for skewed data. *Water Resources Research*, 46(7) :W07519. [11](#)
- Gabriel, K. and Neumann, J. (1962). A markov chain model for daily rainfall occurrence at tel aviv. *Quarterly Journal of the Royal Meteorological Society*, 88(375) :90–95. [12](#)
- Katz, R. W. (1977). Precipitation as a chain-dependent process. *Journal of Applied Meteorology*, 16(7) :671–676. [12](#)
- Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in water resources*, 25(8) :1287–1304. [11](#)
- Kleiber, W., Katz, R. W., and Rajagopalan, B. (2013). Daily minimum and maximum temperature simulation over complex terrain. *The Annals of Applied Statistics*, 7(1) :588–612. [13](#)

- Kotsiantis, S., Kostoulas, A., Lykoudis, S., Argiriou, A., and Menagias, K. (2006). Filling missing temperature values in weather data banks. In *Intelligent Environments, 2006. IE 06. 2nd IET International Conference on*, volume 1, pages 327–334. IET. 6
- Oriani, F., Straubhaar, J., Renard, P., and Mariethoz, G. (2014). Simulation of rainfall time series from different climatic regions using the direct sampling technique. *Hydrology and Earth System Sciences*, 18(8) :3015–3031. 12
- Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L., and Morel, S. (2008). Analysis of near-surface atmospheric variables : Validation of the safran analysis over france. *Journal of applied meteorology and climatology*, 47(1) :92–107. 6
- Rajagopalan, B. and Lall, U. (1999). A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water resources research*, 35(10) :3089–3101. 12
- Richardson, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1) :182–190. 13
- Richardson, C. W. and Wright, D. A. (1984). *WGEN : A model for generating daily weather variables*. US Department of Agriculture, Agricultural Research Service Washington, DC, USA. 13
- Rydén, J. (2014). A statistical analysis of trends for warm and cold spells in uppsala by means of counts. *Geografiska Annaler : Series A, Physical Geography*. 11
- Semenov, M. A. (2007). Simulation of extreme weather events by a stochastic weather generator. *Climate Research*, 35(3) :203. 11
- Semenov, M. A. and Barrow, E. M. (1997). Use of a stochastic weather generator in the development of climate change scenarios. *Climatic change*, 35(4) :397–414. 12
- Semenov, M. A., Brooks, R. J., Barrow, E. M., and Richardson, C. W. (1998). Comparison of the wgen and lars-wg stochastic weather generators for diverse climates. *Climate research*, 10(2) :95–107. 11
- Verdin, A., Rajagopalan, B., Kleiber, W., and Katz, R. W. (2015). Coupled stochastic weather generation using spatial and generalized linear models. *Stochastic Environmental Research and Risk Assessment*, 29(2) :347–356. 13
- Vrac, M. and Naveau, P. (2007). Stochastic downscaling of precipitation : From dry events to heavy rainfalls. *Water resources research*, 43(7). 11, 12
- Wilks, D. (1998). Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210(1) :178–191. 13

- Wilks, D. (1999). Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain. *Agricultural and Forest meteorology*, 96(1) :85–101. [13](#)
- Wilks, D. S. (2010). Use of stochastic weathergenerators for precipitation downscaling. *Wiley Interdisciplinary Reviews : Climate Change*, 1(6) :898–907. [12](#)
- Wilks, D. S. (2012). Stochastic weather generators for climate-change downscaling, part ii : multivariable and spatially coherent multisite downscaling. *Wiley Interdisciplinary Reviews : Climate Change*, 3(3) :267–278. [12](#)
- Wilks, D. S. and Wilby, R. L. (1999). The weather generation game : a review of stochastic weather models. *Progress in Physical Geography*, 23(3) :329–357. [12](#)
- Yiou, P. (2014). Anawege : a weather generator based on analogues of atmospheric circulation. *Geoscientific Model Development*, 7(2) :531–543. [12](#)

Chapitre 2

Rappels sur les champs gaussiens

Sommaire

2.1 Champs gaussiens et modèles de covariance associés	19
2.1.1 Champ aléatoire et fonction de covariance	19
2.1.2 Propriétés des champs gaussiens	20
2.1.3 Modèles de covariance non séparables	25
2.2 Méthodes d'inférence par vraisemblance pour les champs gaussiens . .	30
2.2.1 Vraisemblance complète	30
2.2.2 Vraisemblance effacée	32
2.2.3 Vraisemblance composite	32
2.3 Références	36

2.1 Champs gaussiens et modèles de covariance associés

La classe des processus gaussiens est certainement l'une des plus utilisées parmi les familles de processus stochastiques lorsque l'on cherche à modéliser des données temporelles, spatiales ou spatio-temporelles. La popularité des processus gaussiens est due à quelques propriétés fondamentales que nous allons présenter ensuite dans un cadre multivarié de dimension p . La première est la caractérisation de ses lois fini-dimensionnelles par ses deux premiers moments puisque ce sont des lois normales multivariées. Même si pour de nombreuses données, l'hypothèse de distribution normale n'est pas adaptée, les processus gaussiens sont tout de même employés au travers de constructions complexes (fonctions de transformation, processus latents [Allard, 2012]) pour apporter une structure spatio-temporelle par exemple. De plus, la loi normale possède des propriétés très intéressantes concernant ses distributions conditionnelles, ce qui permet dans un contexte de prédiction de définir facilement le meilleur prédicteur (au sens de l'erreur quadratique) comme une combinaison linéaire de valeurs observées. Enfin comme nous le verrons dans la dernière section de ce chapitre, il est possible d'estimer efficacement les différents paramètres d'une fonction de covariance d'un champ gaussien.

2.1.1 Champ aléatoire et fonction de covariance

Un champ aléatoire $X(\mathbf{k})$ (ou fonction aléatoire) sur $\mathbb{R}^d$ (i.e. $\mathbf{k} \in \mathbb{R}^d, d \in \mathbb{N}$) est une fonction dont les valeurs sont des variables aléatoires quelque soit $\mathbf{k} \in \mathbb{R}^d$, pour $d > 0$. Dans le chapitre précédent, nous avons employé les termes de processus aléatoire et de champ aléatoire. D'un point de vue formel, il n'y a pas de différence, mais dans la pratique le terme de processus est utilisé préférentiellement lorsque $d = 1$ tandis que celui de champ l'est lorsque $d > 1$. Par exemple, les séries temporelles correspondent à $d = 1$, les données spatiales à $d = 2$ ou $d = 3$ suivant les cas. Dans ce manuscrit, nous ne ferons pas nécessairement cette distinction. La loi du champ aléatoire est caractérisée par ses lois fini-dimensionnelles pour toute valeur de n et tout ensemble de n -uplets $\mathbf{k}_i \in \mathbb{R}^d, i = 1, \dots, n$:

$$F_{\mathbf{k}_1, \dots, \mathbf{k}_n}(B_1 \times \dots \times B_n) = \mathbb{P}(\{X(\mathbf{k}_1), \dots, X(\mathbf{k}_n)\} \in B_1 \times \dots \times B_n) \quad (2.1)$$

où $B_1, \dots, B_n$ sont des boréliens de $\mathbb{R}$. Ces lois doivent vérifier les conditions de Kolmogorov de symétrie et de cohérence.

Le moment m_r d'ordre $r \in \mathbb{N}$ d'un champ aléatoire est défini par :

$$m_r(\mathbf{k}) := \mathbb{E}[X(\mathbf{k})^r], \quad \forall \mathbf{k} \in \mathbb{R}^d. \quad (2.2)$$

Si le moment d'ordre 1 est fini, i.e. $m_1(\mathbf{k}) < \infty$ pour tout $\mathbf{k} \in \mathbb{R}^d$, alors on dit que la fonction m_1 est la moyenne du champ aléatoire X . Le champ aléatoire est centré si pour tout $\mathbf{k} \in \mathbb{R}^d, m_1(\mathbf{k}) = 0$.

Un champ aléatoire est dit être à l'ordre 2 si

$$\mathbb{V}[X(\mathbf{k})] := \mathbb{E}[(X(\mathbf{k}) - m_1(\mathbf{k}))^2] < \infty, \quad \forall \mathbf{k} \in \mathbb{R}^d. \quad (2.3)$$

Ceci est équivalent à dire que le moment d'ordre 2 du champ aléatoire existe (i.e. est fini) pour tout $\mathbf{k} \in \mathbb{R}^d$. Il en découle l'existence de la fonction de covariance C définie par $C(\mathbf{k}, \mathbf{k}') = \text{Cov}[X(\mathbf{k}), X(\mathbf{k}')] = \mathbb{E}[X(\mathbf{k})X(\mathbf{k}') - m_1(\mathbf{k})m_1(\mathbf{k}')] = \mathbb{E}[X(\mathbf{k})X(\mathbf{k}') - m_1(\mathbf{k})m_1(\mathbf{k}) - m_1(\mathbf{k})m_1(\mathbf{k}') + m_1(\mathbf{k})m_1(\mathbf{k})]$ pour tout $\mathbf{k}, \mathbf{k}' \in \mathbb{R}^d$, avec $\mathbb{V}[X(\mathbf{k})] = C(\mathbf{k}, \mathbf{k})$. La propriété caractéristique d'une covariance est d'être semi-définie positive, c'est-à-dire que pour tout n , tout n -uplets $\mathbf{k}_1, \dots, \mathbf{k}_n$ de $\mathbb{R}^d$ et pour tout n -uplets $\lambda_1, \dots, \lambda_n$

$$\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(\mathbf{k}_i, \mathbf{k}_j) \geq 0. \quad (2.4)$$

Cette positivité résulte de l'égalité :

$$\mathbb{V}\left[\sum_{i=1}^n \lambda_i X(\mathbf{k}_i)\right] = \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(\mathbf{k}_i, \mathbf{k}_j) \geq 0, \quad (2.5)$$

et qu'une variance ne peut être négative. La covariance est définie positive si l'inégalité est stricte.

2.1.2 Propriétés des champs gaussiens

Dans le cas des champs gaussiens, la spécification des lois fini-dimensionnelles est simple puisque c'est un champ aléatoire dont toutes les lois fini-dimensionnelles sont des lois normales multivariées. Ainsi pour toute valeur de $n \in \mathbb{N}$ et tout ensemble de n -uplets $\mathbf{k}_1, \dots, \mathbf{k}_n \in \mathbb{R}^d$, le vecteur aléatoire gaussien $\mathbf{Z} = (Z_1(\mathbf{k}_1), \dots, Z_p(\mathbf{k}_1), \dots, Z_p(\mathbf{k}_n))^T$ est distribué selon une loi normale multivariée de vecteur $\boldsymbol{\mu}$ d'espérance (de longueur np), moment d'ordre 1,

$$\mathbb{E}[Z_i(\mathbf{k}_\alpha)] = \mu_i(\mathbf{k}_\alpha) \quad (2.6)$$

et de matrice $\boldsymbol{\Sigma}$ de variance-covariance (de dimension $np \times np$), moment d'ordre 2, composée des variances sur sa diagonale (principale) et des éléments de covariance ailleurs

$$\mathbb{V}[Z_i(\mathbf{k}_\alpha)] = \mathbb{E}[Z_i(\mathbf{k}_\alpha) - \mu_i(\mathbf{k}_\alpha)]^2 \quad (2.7)$$

$$\text{Cov}[Z_i(\mathbf{k}_\alpha), Z_j(\mathbf{k}_\beta)] = C_{ij}(\mathbf{k}_\alpha, \mathbf{k}_\beta) \quad (2.8)$$

$$= \mathbb{E}[\{Z_i(\mathbf{k}_\alpha) - \mu_i(\mathbf{k}_\alpha)\}\{Z_j(\mathbf{k}_\beta) - \mu_j(\mathbf{k}_\beta)\}] \quad (2.9)$$

où $\mathbf{k}_\alpha, \mathbf{k}_\beta \in \mathbb{R}^d$, $i, j = 1, \dots, p$ et $\alpha, \beta = 1, \dots, n$. La matrice $\boldsymbol{\Sigma}$ est donc de la forme

$$\boldsymbol{\Sigma} = \begin{bmatrix} \begin{bmatrix} C_{11}(\mathbf{k}_1, \mathbf{k}_1) & \cdots & C_{1p}(\mathbf{k}_1, \mathbf{k}_1) \\ \vdots & \ddots & \vdots \\ C_{p1}(\mathbf{k}_1, \mathbf{k}_1) & \cdots & C_{pp}(\mathbf{k}_1, \mathbf{k}_1) \end{bmatrix} & \cdots & \begin{bmatrix} C_{11}(\mathbf{k}_1, \mathbf{k}_n) & \cdots & C_{1p}(\mathbf{k}_1, \mathbf{k}_n) \\ \vdots & \ddots & \vdots \\ C_{p1}(\mathbf{k}_1, \mathbf{k}_n) & \cdots & C_{pp}(\mathbf{k}_1, \mathbf{k}_n) \end{bmatrix} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} C_{11}(\mathbf{k}_n, \mathbf{k}_1) & \cdots & C_{1p}(\mathbf{k}_n, \mathbf{k}_1) \\ \vdots & \ddots & \vdots \\ C_{p1}(\mathbf{k}_n, \mathbf{k}_1) & \cdots & C_{pp}(\mathbf{k}_n, \mathbf{k}_1) \end{bmatrix} & \cdots & \begin{bmatrix} C_{11}(\mathbf{k}_n, \mathbf{k}_n) & \cdots & C_{1p}(\mathbf{k}_n, \mathbf{k}_n) \\ \vdots & \ddots & \vdots \\ C_{p1}(\mathbf{k}_n, \mathbf{k}_n) & \cdots & C_{pp}(\mathbf{k}_n, \mathbf{k}_n) \end{bmatrix} \end{bmatrix} \quad (2.10)$$

Soit $\mathbf{z}$ une réalisation de $\mathbf{Z}$, la densité de probabilité jointe s'écrit explicitement :

$$f_{\mathbf{Z}}(\mathbf{z}) = (2\pi)^{-np/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu}) \right\}. \quad (2.11)$$

Maintenant présentons une propriété particulière des champs gaussiens qui permet d'obtenir des distributions conditionnelles explicites. Lorsque l'on partitionne un vecteur gaussien en deux sous-ensembles, la distribution conditionnelle de l'un (le "non observé") sachant l'autre (l'"observé") est encore une loi normale multivariée. L'espérance conditionnelle est une combinaison linéaire des observations contrairement à la matrice de covariance qui ne dépend pas des observations. Cette propriété engendre beaucoup d'applications comme la génération de données manquantes ou la prédiction. Pour simplifier la notation, posons $1, \dots, \alpha, \beta, \dots, n \in \mathbb{N}$. Le vecteur aléatoire gaussien $\mathbf{Z} = (Z_1(\mathbf{k}_1), \dots, Z_p(\mathbf{k}_n))^\top$ peut se partitionner en deux vecteurs $\mathbf{Z}_A = (Z_1(\mathbf{k}_1), \dots, Z_p(\mathbf{k}_\alpha))^\top$ et $\mathbf{Z}_B = (Z_1(\mathbf{k}_\beta), \dots, Z_p(\mathbf{k}_n))^\top$ qui ont pour distribution jointe la normale :

$$\begin{bmatrix} \mathbf{Z}_A \\ \mathbf{Z}_B \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \right). \quad (2.12)$$

La distribution conditionnelle de $\mathbf{Z}_A$ sachant $\mathbf{Z}_B = \mathbf{z}_B$ est une loi normale multivariée (de dimension αp) qui a pour espérance et matrice de covariance :

$$\mathbb{E}[\mathbf{Z}_A | \mathbf{Z}_B = \mathbf{z}_B] = \boldsymbol{\mu}_A + \Sigma_{AB} \Sigma_{BB}^{-1} (\mathbf{z}_B - \boldsymbol{\mu}_B), \quad (2.13)$$

$$\text{Cov}[\mathbf{Z}_A | \mathbf{Z}_B = \mathbf{z}_B] = \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA}. \quad (2.14)$$

Le champ gaussien est entièrement caractérisé par ses deux premiers moments. L'espérance peut être choisie arbitrairement mais la fonction de covariance doit être semi-définie positive pour assurer la consistance des lois fini-dimensionnelles. Ainsi pour tout n , pour tout n -uplets $\mathbf{k}_1, \dots, \mathbf{k}_n$ de $\mathbb{R}^d$ et pour tout np -uplets $\lambda_{11}, \dots, \lambda_{np}$, l'inégalité 2.4 doit être vérifiée et dans le cadre multivariée prend la forme suivante :

$$\sum_{i=1}^p \sum_{j=1}^p \sum_{\alpha=1}^n \sum_{\beta=1}^n \lambda_{\alpha i} \lambda_{\beta j} C_{ij}(\mathbf{k}_\alpha, \mathbf{k}_\beta) \geq 0. \quad (2.15)$$

Cette condition est nécessaire et suffisante dans le cas gaussien pour assurer l'existence du champ aléatoire et permet à toutes les matrices Σ , comme définies au dessus, d'être semi-définies positives. En conséquence, une partie du travail de modélisation va se concentrer sur le développement de fonctions de covariance valides, i.e. semi-définies positives. Mais vérifier cette condition n'est pas une tâche aisée et pour faciliter cette opération il est intéressant de restreindre la classe des fonctions de covariance en se basant sur des propriétés de symétrie. A partir de maintenant, nous utiliserons la notation suivante $\mathbf{k} = (\mathbf{s}, t) \in \mathbb{R}^{d+1}$ pour visualiser au mieux l'aspect spatio-temporel.

Un champ gaussien multivarié spatio-temporel (MSTGRF) est dit stationnaire si :

$$\mathbb{E}[Z_i(\mathbf{s}, t)] = \mu_i \quad (2.16)$$

$$\text{Cov}[Z_i(\mathbf{s}, t), Z_j(\mathbf{s} + \mathbf{h}, t + u)] = C_{ij}(\mathbf{h}, u) \quad (2.17)$$

$$= \mathbb{E}[\{Z_i(\mathbf{s}, t) - \mu_i\} \{Z_j(\mathbf{s} + \mathbf{h}, t + u) - \mu_j\}] \quad (2.18)$$

où $(\mathbf{h}, u) \in \mathbb{R}^{d+1}$ et $i, j = 1, \dots, p$. Cette stationnarité est dite faible ou d'ordre 2 dans le cas de champs aléatoires quelconques et elle est équivalente dans le cas gaussien avec la stationnarité au sens strict qui concerne l'invariance des lois fini-dimensionnelles par translation. L'hypothèse de stationnarité peut être trop restrictive et il est possible de construire des champs gaussiens non stationnaires à partir de champs stationnaires. De plus, il est tout à fait possible de définir un champ stationnaire uniquement pour l'espace ou le temps. La matrice Σ de variance covariance du champ stationnaire peut maintenant s'écrire sous une forme simplifiée de 2.10 :

$$\Sigma = \begin{bmatrix} \begin{bmatrix} C_{11}(\mathbf{0}, 0) & \cdots & C_{1p}(\mathbf{0}, 0) \\ \vdots & \ddots & \vdots \\ C_{p1}(\mathbf{0}, 0) & \cdots & C_{pp}(\mathbf{0}, 0) \end{bmatrix} & \cdots & \begin{bmatrix} C_{11}(\mathbf{h}_{\alpha\beta}, u_{\alpha\beta}) & \cdots & C_{1p}(\mathbf{h}_{\alpha\beta}, u_{\alpha\beta}) \\ \vdots & \ddots & \vdots \\ C_{p1}(\mathbf{h}_{\alpha\beta}, u_{\alpha\beta}) & \cdots & C_{pp}(\mathbf{h}_{\alpha\beta}, u_{\alpha\beta}) \end{bmatrix} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} C_{11}(\mathbf{h}_{\beta\alpha}, u_{\beta\alpha}) & \cdots & C_{1p}(\mathbf{h}_{\beta\alpha}, u_{\beta\alpha}) \\ \vdots & \ddots & \vdots \\ C_{p1}(\mathbf{h}_{\beta\alpha}, u_{\beta\alpha}) & \cdots & C_{pp}(\mathbf{h}_{\beta\alpha}, u_{\beta\alpha}) \end{bmatrix} & \cdots & \begin{bmatrix} C_{11}(\mathbf{0}, 0) & \cdots & C_{1p}(\mathbf{0}, 0) \\ \vdots & \ddots & \vdots \\ C_{p1}(\mathbf{0}, 0) & \cdots & C_{pp}(\mathbf{0}, 0) \end{bmatrix} \end{bmatrix} \quad (2.19)$$

où $\mathbf{s}_\alpha - \mathbf{s}_\beta = \mathbf{h}_{\alpha\beta} \in \mathbb{R}^d$ et $t_\alpha - t_\beta = u_{\alpha\beta} \in \mathbb{R}$ avec $\alpha, \beta = 1, \dots, n$. Bien que la matrice Σ de variance covariance soit semi-définie positive, la matrice $\mathbf{C}(\mathbf{h}, u)$ de fonctions de covariance (de dimension $p \times p$), définie par $\mathbf{C}(\mathbf{h}, u) = [C_{ij}(\mathbf{h}, u)]_{i,j=1}^p$ ne l'est pas nécessairement ce qui se traduit par le fait qu'il n'y ait pas d'inéquation du type (Wackernagel [2003], p.136) :

$$C_{ii}(\mathbf{h}, u)C_{jj}(\mathbf{h}, u) \not\geq |C_{ij}(\mathbf{h}, u)|^2. \quad (2.20)$$

Les éléments diagonaux des matrices $[C_{ij}(\mathbf{h}, u)]_{i,j=1}^p$ seront appelés covariances directes et les autres covariances croisées. Le terme fonction de covariance croisée fera référence à $C_{ij}(\mathbf{h}, u)$ quelques soient $(\mathbf{h}, u) \in \mathbb{R}^{d+1}$ et $i, j = 1, \dots, p$.

Même avec l'hypothèse de stationnarité, il reste difficile de prouver la condition 2.15. Les fonctions de covariance stationnaires possèdent une caractérisation dans le domaine spectral plus connue sous le nom de théorème de Bochner (énoncé en 1932 [Bochner, 1955]) et qui a été généralisé au cas multivarié par Cramér [Cramér, 1940]. Enonçons le directement dans le cas où toutes les fonctions de covariance croisée sont intégrables.

Théorème 1. [Cramér, 1940] Les fonctions C_{ij} , continues sur $\mathbb{R}^{d+1}$ forment une matrice $\mathbf{C}$ de fonctions de covariance si et seulement si elles sont de la forme

$$C_{ij}(\mathbf{h}, u) = \int_{\mathbb{R}^d} \int_{\mathbb{R}} e^{i(\boldsymbol{\omega}^\top \mathbf{h} + \tau u)} f_{ij}(\boldsymbol{\omega}, \tau) d\boldsymbol{\omega} d\tau, \quad (\mathbf{h}, u) \in \mathbb{R}^{d+1}, i, j = 1, \dots, p, \quad (2.21)$$

où la matrice des densités spectrales $[f_{ij}(\boldsymbol{\omega}, \tau)]_{i,j=1}^p$ est semi-définie positive pour tout $(\boldsymbol{\omega}, \tau) \in \mathbb{R}^{d+1}$ (Yaglom [1987], p.315 ; Chilès and Delfiner [1999], p.326 ; Wackernagel [2003], p.137).

Ce théorème établit un lien direct entre C_{ij} et sa transformée de Fourier f_{ij} . Il est plus aisé de vérifier la validité de la fonction de covariance croisée à l'aide de sa transformée de Fourier puisqu'il existe une version de l'inégalité 2.20 dans le domaine spectral, et ainsi

$$f_{ii}(\boldsymbol{\omega}, \tau) f_{jj}(\boldsymbol{\omega}, \tau) \geq |f_{ij}(\boldsymbol{\omega}, \tau)|^2 \quad (2.22)$$

pour $(\boldsymbol{\omega}, \tau) \in \mathbb{R}^{d+1}$. Le calcul de l'intégrale sur l'espace $\mathbb{R}^{d+1}$ n'est pas toujours aisé et il existe parfois une forme explicite de la transformée de Fourier calculable plus facilement sur un sous-espace comme $\mathbb{R}^d$. Cette transformée de Fourier partielle pourra être utilisée pour montrer la validité de la fonction de covariance grâce au théorème énoncé dans [Cressie and Huang \[1999\]](#), [Gneiting \[2002\]](#) et généralisé au cas multivarié dans [Schlather \[2010\]](#).

Théorème 2. [Schlather, 2010] La matrice de fonctions $\mathbf{C}(\mathbf{h}, u) = [C_{ij}(\mathbf{h}, u)]_{i,j=1}^p$, définies sur $\mathbb{R}^d \times \mathbb{R}$, est une matrice de fonctions de covariance si et seulement si les fonctions

$$g_{ij}(\boldsymbol{\omega}, u) = (2\pi)^{-d} \int_{\mathbb{R}^d} e^{-i\boldsymbol{\omega}^\top \mathbf{h}} C_{ij}(\mathbf{h}, u) d\mathbf{h}, \quad u \in \mathbb{R}, i, j = 1, \dots, p, \quad (2.23)$$

forment une matrice de fonctions de covariance sur $\mathbb{R}$ pour presque tout $\boldsymbol{\omega} \in \mathbb{R}^d$.

Dans un cadre univarié la fonction de covariance est paire (i.e. $C_{ii}(\mathbf{h}, u) = C_{ii}(-\mathbf{h}, -u)$ avec $i = 1, \dots, p$) ce qui n'est à priori pas le cas pour la fonction de covariance croisée en multivarié. En effet, lorsque l'on procède au changement d'ordre des indices ou du signe du vecteur $(\mathbf{h}, u)$ il n'y a pas de raison d'avoir d'égalité entre les deux fonctions de covariance croisée :

$$C_{ij}(\mathbf{h}, u) \neq C_{ji}(\mathbf{h}, u), \quad i \neq j, \quad (2.24)$$

$$C_{ij}(-\mathbf{h}, -u) \neq C_{ij}(\mathbf{h}, u), \quad i \neq j. \quad (2.25)$$

$$(2.26)$$

En revanche nous obtenons l'égalité suivante lorsque l'on effectue les deux changements en même temps :

$$C_{ij}(\mathbf{h}, u) = C_{ji}(-\mathbf{h}, -u), \quad i \neq j. \quad (2.27)$$

En effet, en reprenant les notations précédentes (supposons l'espérance nulle sans perte de généralité) :

$$\begin{aligned} C_{ji}(-\mathbf{h}, -u) &= \mathbb{E}[Z_j(\mathbf{s}, t) Z_i(\mathbf{s} - \mathbf{h}, t - u)] \\ &= \mathbb{E}[Z_j(\mathbf{s} + \mathbf{h}, t + u) Z_i(\mathbf{s}, t)] \quad (\text{invariant par translation}) \\ &= C_{ij}(\mathbf{h}, u). \end{aligned}$$

Au travers de ces dernières équations, il apparaît un certain de nombres de cas en fonction de l'espace, du temps et des variables. C'est ce que nous allons voir en définissant les notions de séparabilité et de symétrie. Nous nous placerons toujours dans le cadre

d'un champ gaussien stationnaire (sauf mention contraire) multivarié spatio-temporel qui correspond à celui des modèles développés lors de cette thèse. La définition dans un contexte non stationnaire sera très semblable, il suffit de reprendre la notation de l'équation 2.8 avec les définitions qui suivent.

La fonction de covariance croisée d'un champ stationnaire gaussien (MSTGRF) est dite entièrement séparable si

$$C_{ij}(\mathbf{h}, u) = a_{ij} \rho_S(\mathbf{h}) \rho_T(u), \quad (\mathbf{h}, u) \in \mathbb{R}^{d+1}, \quad i, j = 1, \dots, p \quad (2.28)$$

où $A = [a_{ij}]_{i,j=1}^p$ est une matrice réelle définie positive et ρ_S, ρ_T sont respectivement des fonctions de corrélation spatiale sur $\mathbb{R}^d$ et temporelle sur $\mathbb{R}$. Il existe plusieurs niveaux de séparabilité allant d'entièrement séparable à non séparable. Par exemple la séparabilité peut être uniquement entre les variables et l'espace plus le temps et ainsi la fonction de covariance croisée se décompose de la façon suivante :

$$C_{ij}(\mathbf{h}, u) = a_{ij} \rho_{ST}(\mathbf{h}, u), \quad (\mathbf{h}, u) \in \mathbb{R}^{d+1}, \quad i, j = 1, \dots, p \quad (2.29)$$

où $A = [a_{ij}]_{i,j=1}^p$ est toujours une matrice réelle définie positive et ρ_{ST} est une fonction de corrélation spatio-temporelle non séparable. L'équation 2.29 est aussi appelée covariance intrinsèque dans Wackernagel [2003]. Des exemples seront proposés dans la sous section 2.1.3. De même, il est possible d'avoir une séparabilité entre l'espace et les variables plus le temps ou encore le temps et les variables plus l'espace ; cela nécessite respectivement la validité d'un modèle du type $C_{ij}(\mathbf{h})$ ou $C_{ij}(u)$. Ces deux configurations sont très proches puisqu'un modèle de covariance valide dans $\mathbb{R}^d$ le sera nécessairement dans $\mathbb{R}$ mais la réciproque n'est pas toujours vraie. Ces modèles seront respectivement appelés séparable en temps et séparable en espace, et le modèle séparable correspond à l'équation 2.29.

L'hypothèse de séparabilité présente deux gros avantages. Le premier point fort de cette hypothèse est la possibilité de créer simplement des fonctions de covariance croisée valides. En effet, le produit de fonctions semi-définies positives est encore une fonction semi-définie positive ce qui évite d'avoir à utiliser les caractérisations plus lourdes que nous avons présentées précédemment. Deuxièmement les modèles séparables pour des données spatio-temporelles multivariées offrent un gain significatif en terme de temps de calcul. Ces jeux de données sont souvent de grandes tailles et l'estimation par maximum de vraisemblance (voir 2.2) des paramètres de la fonction de covariance croisée nécessite le calcul du déterminant et l'inversion de la matrice Σ (2.19) ce qui est extrêmement coûteux. Or pour les modèles séparables, la matrice Σ peut s'écrire comme le produit de Kronecker de trois matrices de dimensions inférieures

$$\Sigma = A \otimes \Sigma_S \otimes \Sigma_T \quad (2.30)$$

où Σ_S est une matrice de covariance spatiale (de dimension $n_S \times n_S$) et Σ_T une matrice

de covariance temporelle (de dimension $n_T \times n_T$) avec $n = n_S \times n_T$. Dans ce cas,

$$\Sigma^{-1} = A^{-1} \otimes \Sigma_S^{-1} \otimes \Sigma_T^{-1} \quad (2.31)$$

$$|\Sigma| = |A|^n |\Sigma_S|^{pn_T} |\Sigma_T|^{pn_S} \quad (2.32)$$

ce qui économise énormément de temps de calcul. Le contrecoup de cette simplicité est que l'hypothèse de séparabilité n'est pas toujours adaptée pour rendre compte de la complexité des interactions (directes et croisées) entre variables, espace et temps, en particulier pour des données météorologiques. En effet, la définition 2.29 implique comme contrainte que toutes les variables aient la même dynamique spatio-temporelle. Ainsi si l'on observe le ratio pour deux vecteurs spatio-temporels différents

$$\frac{C_{ij}(\mathbf{h}, u)}{C_{ij}(\mathbf{h}', u')} = \frac{\rho_S(\mathbf{h}) \rho_T(u)}{\rho_S(\mathbf{h}') \rho_T(u')}, \quad \mathbf{h}, \mathbf{h}' \in \mathbb{R}^d, u, u' \in \mathbb{R}, \quad (2.33)$$

le terme de droite ne dépend ni de i ni de j . Soient $\mathbf{s}, \mathbf{s}' \in \mathbb{R}^d$ et $t, t' \in \mathbb{R}$ avec $\mathbf{s} - \mathbf{s}' = \mathbf{h}$ et $t - t' = u$, l'équation 2.29 est équivalente à :

$$C_{ij}(\mathbf{h}, u) = \text{Cov}[Z_i(\mathbf{s}, t)Z_j(\mathbf{s}, t)] \times \text{Cov}[Z_j(\mathbf{s}, t)Z_j(\mathbf{s}', t)] \times \text{Cov}[Z_j(\mathbf{s}', t)Z_j(\mathbf{s}', t')] \quad (2.34)$$

$$= \text{Cov}[Z_i(\mathbf{s}, t)Z_i(\mathbf{s}, t')] \times \text{Cov}[Z_i(\mathbf{s}, t')Z_i(\mathbf{s}', t')] \times \text{Cov}[Z_i(\mathbf{s}', t')Z_j(\mathbf{s}', t')] \quad (2.35)$$

ce qui signifie que $Z_i(\mathbf{s}, t)$ et $Z_j(\mathbf{s}', t')$ sont indépendants conditionnellement à $Z_j(\mathbf{s}, t)$ et $Z_j(\mathbf{s}', t)$ ou $Z_i(\mathbf{s}, t')$ et $Z_i(\mathbf{s}', t')$.

Enfin, La fonction de covariance croisée d'un champ stationnaire gaussien (MSTGRF) est dite entièrement symétrique si

$$\begin{aligned} C_{ij}(\mathbf{h}, u) &= C_{ij}(-\mathbf{h}, u), \\ &= C_{ij}(\mathbf{h}, -u), \\ &= C_{ij}(-\mathbf{h}, -u), \quad (\mathbf{h}, u) \in \mathbb{R}^{d+1}, i, j = 1, \dots, p. \end{aligned} \quad (2.36)$$

La dernière ligne équivaut à $C_{ij}(\mathbf{h}, u) = C_{ji}(\mathbf{h}, u)$ parfois nommée hypothèse de symétrie. Une fonction de covariance (entièrement) séparable est nécessairement entièrement symétrique mais la réciproque n'est pas toujours vraie. En effet, les termes de covariance (ou corrélation) de 2.28 et 2.29 sont univariés donc paires. La figure 2.1, élaborée à partir des figures dans Gneiting et al. [2007] et De Iaco et al. [2013], met en évidence les liens entre les différentes propriétés.

2.1.3 Modèles de covariance non séparables

Dans la suite lorsque nous parlerons de champs gaussiens, les propriétés de stationnarité, symétrie, séparabilité seront attribuées de façon équivalente au champ aléatoire gaussien ou à la fonction de covariance qui le caractérise. Sauf si cela est mentionné, nous nous plaçons dans le cadre de champs gaussiens stationnaires. Cette sous section va

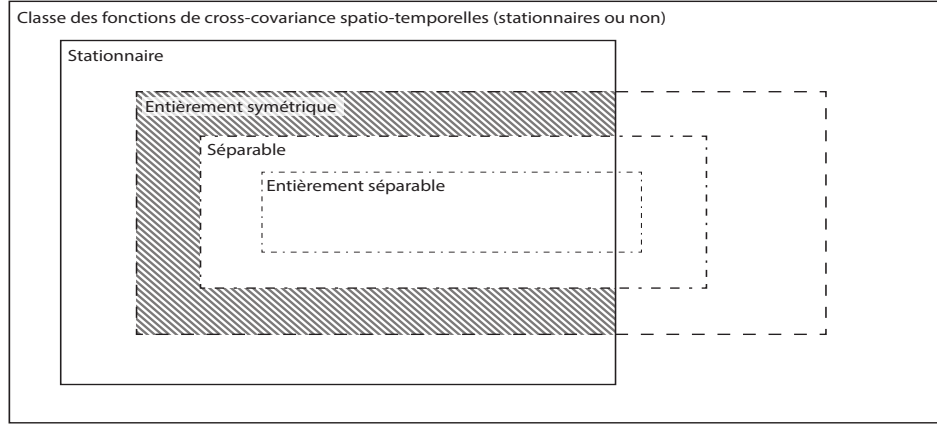


FIGURE 2.1 – Illustration des relations entre les différentes propriétés des fonctions de covariance stationnaires ou non.

présenter deux exemples de fonctions de covariance valides très utilisées en séries temporelles ou en géostatistique et qui seront à la base de la construction de notre modèle. Ces fonctions sont isotropes, c'est à dire qu'en plus d'être invariantes par translation, elles le sont aussi par rotation et réflexion. En conséquence elles n'ont pas de direction privilégiée et nous utiliserons la norme euclidienne comme distance. L'isotropie implique la symétrie et la classe de fonctions de covariance croisée que nous proposons le sera aussi (zone hachurée dans la figure 2.1). Cette hypothèse peut être pénalisante car pas forcément vérifiée par les données mais il faut voir la construction de fonctions de covariance croisée valides comme la première étape pour la construction de modèles plus complexes puisqu'il en existe assez peu.

La classe de Matérn est une classe de fonctions de covariance très riche, extrêmement utilisée en modélisation spatiale (définie sur $\mathbb{R}^d$) ou temporelle (et donc valide sur $\mathbb{R}$ aussi). Whittle [1963] a démontré que le champ Y stationnaire isotrope avec une fonction de covariance de Matérn est solution de l'équation différentielle stochastique suivante

$$\left\{ \frac{\partial^2}{\partial s_1^2} + \dots + \frac{\partial^2}{\partial s_d^2} - r^2 \right\}^{\kappa/2} Y(\mathbf{s}) = \delta(\mathbf{s}), \quad (2.37)$$

où $r > 0$, $\mathbf{s} = (s_1, \dots, s_d)^\top \in \mathbb{R}^d$, $\delta(\cdot)$ est un bruit blanc dans $\mathbb{R}^d$ et où l'opérateur linéaire du côté gauche de l'équation est un laplacien fractionnaire (de fraction $\kappa/2$). Le lien entre

κ et le paramètre ν de la covariance de Matérn est donné dans [Cressie and Wikle \[2011\]](#) (page 302) et [Lindgren et al. \[2011\]](#) par $\kappa = \nu + d/2$. La classe de fonctions de covariance de Matérn s'écrit

$$\mathcal{M}(\mathbf{h}|r, \nu) = \frac{2^{1-\nu}}{\Gamma(\nu)} (r \|\mathbf{h}\|)^\nu \mathcal{K}_\nu(r \|\mathbf{h}\|), \quad \mathbf{h} \in \mathbb{R}^d, \quad (2.38)$$

où $r > 0$ est le paramètre d'échelle ($1/r$ est souvent appelé la portée en géostatistique) et $\mathcal{K}_\nu$ correspond à la fonction de Bessel modifiée de second genre d'ordre ν [[Abramowitz and Stegun, 1972](#)]. Le dernier paramètre est le paramètre de régularité $\nu > 0$ qui permet de faire le lien directement entre la fonction de covariance et la régularité du champ aléatoire. En effet suivant la valeur de ν , plusieurs cas sont à présenter. Premièrement pour $\nu < 1/2$, le champ gaussien n'est pas continu en moyenne quadratique ce qui est équivalent au fait que sa fonction de covariance ne soit pas continue à l'origine et implique que le champ soit très rugueux. Le cas $\nu = 1/2$ correspond à une fonction de covariance particulièrement utilisée puisqu'il s'agit du modèle exponentiel $C(\mathbf{h}) = \exp(-r \|\mathbf{h}\|)$. Le champ associé est maintenant continu en moyenne quadratique mais non différentiable en moyenne quadratique à l'origine. Pour le cas plus général $\nu = k + 1/2$, où k est un entier positif, la covariance de Matérn devient le produit d'une covariance exponentielle avec un polynôme de degré k à coefficients positifs. Le champ est quant à lui k fois différentiable puisqu'un champ gaussien est k fois différentiable en moyenne quadratique si et seulement si $\nu \geq k$ [[Banerjee and Gelfand, 2003](#)]. Enfin plus la valeur de ν est grande et plus le champ associé est lisse jusqu'au cas limite où $\nu \rightarrow \infty$ qui conduit à la fonction de covariance gaussienne $C(\mathbf{h}) = \exp(-r^2 \|\mathbf{h}\|)$ et à un champ infiniment différentiable en moyenne quadratique. La flexibilité de la covariance de Matérn, liée au paramètre de régularité ν est une des raisons de la popularité de cette famille de fonctions.

Une autre raison tient dans l'écriture explicite de la transformée de Fourier de la covariance de Matérn

$$f(\boldsymbol{\omega}; r, \nu) = \frac{\Gamma(\nu + d/2) r^{2\nu}}{\Gamma(\nu) \pi^{d/2}} (r^2 + \|\boldsymbol{\omega}\|^2)^{-\nu - d/2}, \quad \boldsymbol{\omega} \in \mathbb{R}^d. \quad (2.39)$$

Cette transformée de Fourier est une des pierres angulaires des démonstrations de la validité des fonctions de covariances multivariées de Matérn proposées dans [Gneiting et al. \[2010\]](#) et [Apanasovich et al. \[2012\]](#). Ces fonctions de covariance croisée sont de la forme

$$C_{ij}(\mathbf{h}) = \sigma_i \sigma_j \rho_{ij} \mathcal{M}(\mathbf{h}; r_{ij}, \nu_{ij}), \quad \mathbf{h} \in \mathbb{R}^d. \quad (2.40)$$

Avant de discuter des paramétrisations de $\{r_{ij}, \nu_{ij}, \rho_{ij}\}_{i,j=1,\dots,p}$ proposées dans ces articles, il est important de noter que ces modèles sont non séparables puisque dans le cas contraire l'équation 2.40 se serait écrite ainsi

$$C_{ij}(\mathbf{h}) = \sigma_i \sigma_j \rho_{ij} \mathcal{M}(\mathbf{h}; r, \nu), \quad \mathbf{h} \in \mathbb{R}^d, \quad (2.41)$$

où $[\rho_{ij}]_{i,j=1}^p$ serait une matrice de corrélation. Contrairement au modèle séparable, les modèles non séparables permettent à chaque variable de posséder sa propre dynamique

spatiale tout en permettant des corrélations croisées entre variables. Les premiers modèles proposés par [Gneiting et al. \[2010\]](#) donnent des conditions nécessaires et suffisantes pour $\{r_{ij}, v_{ij}, \rho_{ij}\}_{i,j=1,\dots,p}$ dans le cas bivarié, i.e. $p = 2$, et présentent des conditions suffisantes pour $p \geq 2$ qui sont :

$$\begin{aligned} r_{ij} &= r, \quad 1 \leq i, j \leq p, \\ v_{ij} &= (v_i + v_j)/2, \quad 1 \leq i, j \leq p, \\ \rho_{ij} &= \beta_{ij} \frac{\Gamma(v_i + d/2)^{1/2}}{\Gamma(v_i)^{1/2}} \frac{\Gamma(v_j + d/2)^{1/2}}{\Gamma(v_j)^{1/2}} \frac{\Gamma(v_{ij})}{\Gamma(v_{ij} + d/2)} \quad 1 \leq i, j \leq p, \end{aligned} \quad (2.42)$$

où $[\beta_{ij}]_{i,j=1}^p$ est une matrice de corrélation. La forme de ρ_{ij} paraît relativement compliquée mais ces différents termes sont nécessaires pour permettre à ρ_{ij} d'être égal à 1 lorsque $i = j$ et simplifier l'écriture de la densité spectrale (ou transformée de Fourier) en supprimant le terme $\Gamma((v_i + v_j)/2 + d/2)$. La paramétrisation proposée par [Apanasovich et al. \[2012\]](#) est assez riche et autorise au paramètre d'échelle r de varier pour chaque composante lorsque $p \geq 2$. Ici, nous n'avons retenu qu'une forme plus simplifiée des diverses paramétrisations présentées. Ainsi la fonction de covariance croisée 2.40 est valide si :

$$\begin{aligned} r_{ij} &= \{(r_i^2 + r_j^2)/2\}, \quad 1 \leq i, j \leq p, \\ v_{ij} &= (v_i + v_j)/2, \quad 1 \leq i, j \leq p, \\ \rho_{ij} &= \beta_{ij} \frac{\Gamma(v_{ij})}{\Gamma(v_i)^{1/2} \Gamma(v_j)^{1/2}} \frac{r_{ij}^{v_{ij}}}{(r_i^{v_i} r_j^{v_j})^{1/2}} \quad 1 \leq i, j \leq p, \end{aligned} \quad (2.43)$$

où $[\beta_{ij}]_{i,j=1}^p$ est une matrice de corrélation. Comme pour le modèle précédent, ρ_{ij} est défini de manière à être égal à 1 pour $i = j$ et ainsi obtenir que pour chaque variable, $C_i(\mathbf{h}) = \sigma_i^2 \mathcal{M}(\mathbf{h}; r_i, v_j)$. Nous avons retenu ce cas particulier de paramétrisation car pour les autres cas proposés (auxquels appartient 2.43), un nombre de paramètres importants s'ajoutait lorsque le nombre de variables p augmentait. Or pour nos applications avec un générateur stochastique de climat, la valeur de p sera de 3 à 5 et des paramètres temporels viendront s'ajouter. Enfin une dernière remarque commune aux deux modèles, la condition $v_{ij} = (v_i + v_j)/2$ est le cas limite de la condition suffisante $v_{ij} \geq (v_i + v_j)/2$ et qui est même nécessaire pour $p = 2$. En d'autres termes, la régularité de la covariance croisée est au moins égale à la moyenne des régularités des deux processus marginaux.

La famille de Matérn dans le cas multivarié est un exemple de modèles non séparables entre l'espace et les variables. D'un point de vue mathématique, ces modèles seraient tous aussi valides dans $\mathbb{R}^{d+1}$ en considérant le temps comme une dimension supplémentaire de l'espace. Cependant d'un point de vue physique, cet argument est plus difficilement acceptable. La classe de fonctions de covariance entièrement symétriques non séparables

développées dans Gneiting [2002] apporte une réponse à ce problème en différenciant l'espace du temps dans le modèle.

Avant de présenter plus en détail la classe de covariance de Gneiting, rappelons quelques définitions mathématiques. Une fonction φ continue définie sur $\mathbb{R}_+$ est dite complètement monotone si elle est indéfiniment dérivable et si $(-1)^n \varphi^{(n)}(t) \geq 0$ où $\varphi^{(n)}$ est la dérivée n -ième. Des exemples de fonctions complètement monotones sont présentés dans le tableau 2.1. Celles-ci ont été standardisées, c'est-à-dire que $\varphi(0) = 1$, et les démonstrations de leur monotonie complète sont disponibles dans Gneiting [1997]. Ces fonctions interviennent dans deux théorèmes importants que nous allons énoncer et qui mettent en évidence le lien très fort avec les fonctions de covariance isotropiques.

Le premier théorème formulé par S. Bernstein (1928) (Feller [1966], p.439) est le point de départ de nombreux travaux et notamment de la démonstration de la validité des fonctions de covariance de Gneiting.

Théorème 3.[Bernstein, 1928] Une fonction φ définie sur $\mathbb{R}_+$ est la transformée de Laplace d'une probabilité si et seulement si elle est complètement monotone et $\varphi(0) = 1$. Le même théorème est parfois présenté sous une autre version. Une fonction φ définie sur $\mathbb{R}_+$ est complètement monotone si et seulement si elle est de la forme

$$\varphi(t) = \int_0^\infty \exp(-rt) dF(r), \quad (2.44)$$

où F est une mesure non décroissante.

Le second théorème, appelé théorème de Schoenberg [Schoenberg, 1938], établit directement l'équivalence entre les fonctions de covariance isotropiques et les fonctions complètement monotones.

Théorème 4.[Schoenberg, 1938] La fonction isotropique $C(\mathbf{h}) = \varphi(\|\mathbf{h}\|^2)$, définie sur $\mathbb{R}^d$, est une fonction de covariance spatiale sur $\mathbb{R}^d$ si et seulement si la fonction $\varphi(t)$, $t \geq 0$, est complètement monotone.

Un regard avisé aura certainement reconnu dans le tableau 2.1 trois grandes classes de fonctions de covariance isotropiques en remplaçant t par $\|\mathbf{h}\|$. La première est la famille des covariances exponentielles puissance. Le cas $\gamma = 1/2$ correspond au cas exponentiel classique tandis que $\gamma = 1$ renvoie à la covariance gaussienne. La seconde ligne du tableau n'est plus à présenter puisque nous venons de lui consacrer une partie entière, il s'agit bien entendu de la famille de Matérn. Enfin la dernière conduit à une famille de covariance dont nous n'avons pas encore parlé : la classe de Cauchy [Gneiting and Schlather, 2004].

Maintenant que les définitions sont posées, nous pouvons introduire les conditions suffisantes exposées dans Gneiting [2002] pour construire une classe de fonctions de covariance spatio-temporelles non séparables. Soit la fonction $\varphi(t)$, $t \geq 0$, complètement monotone et soit la fonction $\psi(t)$, $t \geq 0$, positive avec une dérivée complètement mono-

Fonction	Paramètres
$\varphi(t) = \exp(-ct^\gamma)$	$c > 0, 0 < \gamma \leq 1$
$\varphi(t) = (2^{v-1} \Gamma(v))^{-1} (ct^{1/2})^v \mathcal{K}_v(ct^{1/2})$	$c > 0, v > 0$
$\varphi(t) = (1 + ct^\gamma)^{-v}$	$c > 0, 0 < \gamma \leq 1, v > 0$

 TABLE 2.1 – Exemples de fonctions complètement monotones $\varphi(t)$, $t \geq 0$.

Fonction	Paramètres
$\psi(t) = (\alpha t^a + 1)^b$	$\alpha > 0, 0 < a \leq 1, 0 \leq b \leq 1$
$\psi(t) = \log(\alpha t^a + 1) / \log(\beta)$	$\alpha > 0, \beta > 1, 0 < a \leq 1$
$\psi(t) = (\alpha t^a + \beta) / (\beta(\alpha t^a + 1))$	$\alpha > 0, 0 < \beta \leq 1, 0 < a \leq 1$

 TABLE 2.2 – Exemples de fonctions positives $\psi(t)$, $t \geq 0$ avec une dérivée complètement monotone.

tone et soit $\sigma^2 > 0$ alors

$$C(\mathbf{h}, u) = \frac{\sigma^2}{\psi(|u|^2)^{d/2}} \varphi\left(\frac{\|\mathbf{h}\|^2}{\psi(|u|^2)}\right), \quad (\mathbf{h}, u) \in \mathbb{R}^{d+1}, \quad (2.45)$$

est une fonction de covariance spatio-temporelle valide. La preuve de la validité de ce modèle est basé sur l'utilisation combinée des théorèmes 2, 3 et 4. Le théorème 3 permet de remplacer φ par son écriture en transformée de Laplace. Ensuite le calcul de la transformée g de Fourier partielle est effectué par rapport à $\mathbb{R}^d$ et donc d'après le théorème 2, il faut montrer que $g(\omega, u)$ est une fonction de covariance temporelle. Le théorème 4 est utilisé dans l'argumentaire de cette dernière partie de la preuve.

Le tableau 2.2 présente plusieurs exemples de fonctions positives avec une dérivée complètement monotone mais nous n'utiliserons que la première entrée, qui peut être vue comme l'inverse d'une covariance de Cauchy, pour des raisons pratiques expliquées dans Gneiting [2002] et au chapitre suivant. Il est à noter que dans l'équation 2.45, un ratio entre l'espace et une transformée du temps apparaît, ce qui peut être vu d'un point de vue physique comme une vitesse. Enfin, le modèle a été présenté sous sa version spatio-temporelle mais il peut être étendu à une version plus générale travaillant dans $\mathbb{R}^l \times \mathbb{R}^m$, avec $l, m \in \mathbb{N}^*$. Ce qui veut dire que les rôles joués par l'espace et le temps peuvent être inversés.

2.2 Méthodes d'inférence par vraisemblance pour les champs gaussiens

2.2.1 Vraisemblance complète

Une fois l'étape de construction de modèles de covariance achevée, il est nécessaire de procéder à l'estimation des différents paramètres. Dans le cadre des champs gaus-

siens (univariés ou multivariés), l'inférence par maximum de vraisemblance (*maximum likelihood* en anglais) est généralement considérée comme l'une des meilleures méthodes pour estimer les paramètres des fonctions de covariance. En statistique, la vraisemblance mesure l'adéquation entre la distribution observée sur un échantillon aléatoire de taille finie (np) et une loi de probabilité supposée décrire une réalité sur la population dont l'échantillon est issu. Dans le cadre gaussien, la densité jointe f_Z a été définie par l'équation 2.11, et l'on peut écrire la fonction de vraisemblance $f_Z(\mathbf{z}; \theta)$ comme une fonction des paramètres θ , appartenant à l'espace des paramètres $\Theta \subset \mathbb{R}^q$, pour les observations $\mathbf{z} = (z_1(\mathbf{s}_1, t_1), \dots, z_p(\mathbf{s}_n, t_n))$. On peut avoir recours à l'utilisation d'un modèle paramétrique pour les deux premiers moments mais dans nos travaux le champ gaussien sera toujours d'espérance nulle, et l'espace Θ fera référence aux paramètres de la fonction de covariance croisée $C_{ij}(\cdot, \cdot; \theta)$. Pour alléger les notations, l'indexation par θ sera souvent omise. Il ne reste plus qu'à maximiser la fonction de vraisemblance - de manière explicite en annulant les dérivées partielles ou de manière computationnelle à l'aide d'algorithmes numériques - pour obtenir les estimateurs θ_{FL} de θ par maximum de vraisemblance. Les initiales FL font référence à *full likelihood* ou vraisemblance complète pour la différencier des vraisemblances composites que nous définirons plus loin. Dans la pratique, on utilise la log-vraisemblance, ce qui permet de remplacer des produits par des sommes et ne change en rien les estimateurs puisque la fonction logarithme est croissante. Ainsi, l'estimateur du maximum de vraisemblance est

$$\hat{\theta}_{\text{FL}} = \arg \max_{\theta \in \Theta} \log f_Z(\mathbf{z}; \theta). \quad (2.46)$$

Sous des hypothèses de régularité, l'estimateur du maximum de vraisemblance admet pour loi asymptotique

$$\sqrt{n}(\hat{\theta}_{\text{FL}} - \theta_v) \xrightarrow{d} \mathcal{N}_q(0, \mathbf{I}^{-1}(\theta_v)), \quad (2.47)$$

où θ_v est le vecteur des vrais paramètres, avec $\theta_v \in \Theta$, et $\mathbf{I}(\theta_v) = -\mathbb{E}[\nabla^2 \log f_Z(\mathbf{z}; \theta_v)]$ la matrice d'information de Fisher.

La maximisation de la log-vraisemblance nécessite le calcul du déterminant et de l'inverse de la matrice $\Sigma(\theta)$ de variance-covariance (de dimension $np \times np$). Les algorithmes communément utilisés, comme la décomposition de Cholesky, ont un coût de calcul de l'ordre de $\mathcal{O}((np)^3)$ ce qui les rend inutilisables quand n est trop grand (i.e. np supérieur à quelques milliers).

Même si la vraisemblance complète reste la référence pour l'estimation dans le cadre de champs gaussiens, l'utilisation de ces derniers pour des jeux de données de grande dimension a nécessité le développement de nouvelles méthodes d'inférence. Les approches par vraisemblance composite et *tapered* sont des méthodes adaptées à la problématique des champs gaussiens spatio-temporels multivariés. Les développements théoriques dans ce sens sont extrêmement récents.

2.2.2 Vraisemblance effacée

L'approche par *tapered likelihood* ou vraisemblance effacée, proposée par [Kaufman et al. \[2008\]](#), consiste à fixer certains éléments de la matrice $\Sigma(\theta)$ de variance-covariance à zéro pour obtenir une matrice creuse (pour laquelle des algorithmes adaptés et efficaces existent pour calculer le déterminant et l'inverse). Les éléments de la matrice $\Sigma(\theta)$ fixés à zéro sont ceux correspondant aux paires d'observations trop distantes et dont la corrélation est souvent proche de zéro. Autrement dit, ces paires d'observations sont considérées comme indépendantes (cadre gaussien). La perte d'information reste minimale tandis que le gain computationnel est important. Cependant, la nouvelle matrice $\Sigma_T(\theta)$ ainsi créée, et appelée *tapered matrix* ou matrice effacée, doit toujours être semi-définie positive. Pour se faire, on effectue une multiplication terme à terme (produit de Schur) entre la matrice $\Sigma(\theta)$ et une matrice semi-définie positive $\mathbf{T}(d)$ obtenue à l'aide d'une fonction de covariance $r(\mathbf{h}; d)$ isotropique à support compact, appelée *taper function* ou fonction effaceuse, tels que

$$\Sigma_T(\theta) = \Sigma(\theta) \circ \mathbf{T}(d), \quad (2.48)$$

$$r(\mathbf{h}; d) = 0, \quad \text{si } \|\mathbf{h}\| > d > 0, \quad (2.49)$$

où $\circ$ correspond au produit de Schur, et $\mathbf{h} \in \mathbb{R}^d$. L'extension au cas multivarié est proposée dans [Bevilacqua et al. \[2015\]](#) et la fonction de covariance $r(\mathbf{h}; d)$ est remplacée par une matrice $\mathbf{R}$ (de dimension $p \times p$) de fonctions de covariance à support compact. La construction de la matrice $\mathbf{R}$ est l'objet de recherche [\[Daley et al., 2014\]](#), et à l'instar de la section précédente, la propriété de séparabilité permet d'obtenir des modèles multivariés de fonctions effaceuses à partir de fonctions effaceuses univariées. Ces fonctions multivariées peuvent aussi être construites à partir du modèle linéaire de co-régionalisation [\[Wackernagel, 2003\]](#). [Daley et al. \[2014\]](#) fournissent des conditions suffisantes pour la construction de modèles non séparables pour les champs aléatoires multivariés.

La vraisemblance effacée est une approximation de l'équation 2.11 puisque la matrice $\Sigma(\theta)$ est remplacée par la matrice $\Sigma_T(\theta)$ et la log-vraisemblance s'écrit

$$\log f_Z(\mathbf{z}; \theta) = -\frac{1}{2} \log |\Sigma_T(\theta)| - \frac{1}{2} \mathbf{z}^\top ([\Sigma_T(\theta)]^{-1} \circ \mathbf{T}(\mathbf{d})) \mathbf{z}, \quad (2.50)$$

où le vecteur $\mathbf{d}$ est fixé et $\mathbf{T}(\mathbf{d})$ est la matrice (de dimension $np \times np$) associée à $\mathbf{R}(\mathbf{d})$. L'inverse de la matrice $\Sigma_T(\theta)$ a été multipliée (terme à terme) par la matrice $\mathbf{T}(\mathbf{d})$ pour obtenir de nouveau une matrice creuse et faciliter le calcul. De plus, l'estimateur ainsi obtenu est consistant et possède de bonnes propriétés asymptotiques similaires à celles de l'estimateur du maximum de vraisemblance composite (CL).

2.2.3 Vraisemblance composite

Une autre alternative basée sur la vraisemblance a été développée pour estimer les paramètres de la fonction de covariance, ou covariance croisée dans le cas multivarié,

lorsque le nombre de données devient trop important. La vraisemblance composite, qui consiste à multiplier plusieurs "plus petites" vraisemblances (ou additionner des log-vraisemblances), est une méthode d'estimation adaptée à cette problématique de grands jeux de données.

Soit $\{\mathcal{A}_1, \dots, \mathcal{A}_K\}$ un ensemble d'événements marginaux ou conditionnels associés aux vraisemblances $\mathcal{L}_k(\theta; \mathbf{z}) \propto f(\mathbf{z} \in \mathcal{A}_k; \theta)$, $k = 1, \dots, K$, Lindsay [1988] définit comme vraisemblance composite (CL) le produit pondéré

$$\text{CL}(\theta; \mathbf{z}) = \prod_{k=1}^K \mathcal{L}_k(\theta; \mathbf{z})^{w_k}, \quad (2.51)$$

où les w_k sont des poids positifs. L'estimateur $\hat{\theta}_{\text{CL}}$ de θ par maximum de vraisemblance composite est donc

$$\hat{\theta}_{\text{CL}} = \arg \max_{\theta \in \Theta} \text{CL}(\theta; \mathbf{z}). \quad (2.52)$$

Cette définition est très générale et inclut beaucoup de différentes versions de vraisemblances composites en fonction des sous ensembles considérés. Le but est bien évidemment que les vraisemblances des sous ensembles choisis soient facilement calculables. Une présentation complète des différents types de vraisemblances composites est disponible dans Varin et al. [2011]. Dans la suite, nous allons uniquement nous intéresser aux sous ensembles impliquant une paire d'observations. Dans le cadre gaussien, trois versions de la vraisemblance par paires peuvent être facilement définies :

- soit $\mathcal{A}_k = (Z_i(\mathbf{s}_\alpha, t_\alpha), Z_j(\mathbf{s}_\beta, t_\beta))^\top$ avec $i, j = 1, \dots, p$ et $\alpha, \beta = 1, \dots, n$. La vraisemblance composite construite à partir de ces sous ensembles est appelée vraisemblance gaussienne marginale par paires (*pairwise marginal Gaussian likelihood*) ;
- soit $\mathcal{A}_k = Z_i(\mathbf{s}_\alpha, t_\alpha) | Z_j(\mathbf{s}_\beta, t_\beta)$ avec $i, j = 1, \dots, p$ et $\alpha, \beta = 1, \dots, n$. La vraisemblance composite construite à partir de ces sous ensembles est appelée vraisemblance gaussienne conditionnelle par paires (*pairwise conditional Gaussian likelihood*) ;
- soit $\mathcal{A}_k = Z_i(\mathbf{s}_\alpha, t_\alpha) - Z_j(\mathbf{s}_\beta, t_\beta)$ avec $i, j = 1, \dots, p$ et $\alpha, \beta = 1, \dots, n$. La vraisemblance composite construite à partir de ces sous ensembles est appelée vraisemblance gaussienne par paires par différence (*pairwise difference Gaussian likelihood*).

Des simulations numériques ont été effectuées dans Bevilacqua and Gaetan [2014] pour comparer ces trois méthodes avec un champ gaussien spatial univarié. Il en ressort que (dans le cas pondéré) les vraisemblances marginales et conditionnelles offrent de meilleures performances d'estimation que la vraisemblance par différence. De plus, en supposant que l'espérance des accroissements croisés soit nulle

$$\mathbb{E} [Z_i(\mathbf{s}_\alpha, t_\alpha) - Z_j(\mathbf{s}_\beta, t_\beta)] = 0, \quad (2.53)$$

la distribution de $Z_i(\mathbf{s}_\alpha, t_\alpha) - Z_j(\mathbf{s}_\beta, t_\beta)$ est une loi normale

$$Z_i(\mathbf{s}_\alpha, t_\alpha) - Z_j(\mathbf{s}_\beta, t_\beta) \sim \mathcal{N}(0, 2\pi_{ij}(\mathbf{s}_\alpha - \mathbf{s}_\beta, t_\alpha - t_\beta)), \quad (2.54)$$

$$\pi_{ij}(\mathbf{s}_\alpha - \mathbf{s}_\beta, t_\alpha - t_\beta) = \frac{1}{2} \mathbb{E} \left[\left(Z_i(\mathbf{s}_\alpha, t_\alpha) - Z_j(\mathbf{s}_\beta, t_\beta) \right)^2 \right], \quad (2.55)$$

où $\pi_{ij}(\mathbf{s}_\alpha - \mathbf{s}_\beta, t_\alpha - t_\beta)$ est le pseudo variogramme croisé [Wackernagel, 2003]. Le pseudo variogramme croisé prend uniquement des valeurs positives contrairement aux fonctions de covariance croisée ce qui le rend difficilement utilisable lorsqu'il faut modéliser des variables présentant des corrélations négatives. Toutes ces raisons font que nous n'avons pas retenu la vraisemblance par paires par différence dans cette thèse. Enfin, la vraisemblance conditionnelle gaussienne par paires peut s'écrire en fonction de celle marginale lorsque les poids sont symétriques. Tout le travail présenté dans ce manuscrit a été réalisé en utilisant la vraisemblance gaussienne marginale par paires.

Considérons le vecteur $(Z_i(\mathbf{s}_\alpha, t_\alpha), Z_j(\mathbf{s}_\beta, t_\beta))^T$ avec $i, j = 1, \dots, p$ et $\alpha, \beta = 1, \dots, n$. La log-vraisemblance de cette paire s'écrit

$$l(i, j, \mathbf{s}_\alpha, \mathbf{s}_\beta, t_\alpha, t_\beta; \theta) = -\frac{1}{2} \left\{ \log \Delta_{ij, \alpha\beta} + \frac{A_{ij, \alpha\beta}}{\Delta_{ij, \alpha\beta}} \right\}, \quad (2.56)$$

où $\Delta_{ij, \alpha\beta} = \sigma_i^2 \sigma_j^2 - C_{ij}(\mathbf{h}, u)^2$ et $A_{ij, \alpha\beta} = \sigma_j^2 Z_i(\mathbf{s}_\alpha, t_\alpha)^2 - C_{ij}(\mathbf{h}, u) Z_i(\mathbf{s}_\alpha, t_\alpha) Z_j(\mathbf{s}_\beta, t_\beta) + \sigma_i^2 Z_j(\mathbf{s}_\beta, t_\beta)^2$, avec $\mathbf{h} = \|\mathbf{s}_\alpha - \mathbf{s}_\beta\|$, $u = |t_\alpha - t_\beta|$, $\sigma_i = C_{ii}(\mathbf{0}, 0)$. A partir de ces "sous" log-vraisemblances, ou briques élémentaires, nous pouvons construire la log-vraisemblance marginale gaussienne par paires pondérées comme la somme pondérée de ces briques :

$$\text{wpl}(\theta) = \sum_{(i, j, \alpha, \beta) \in \Lambda} l(i, j, \mathbf{s}_\alpha, \mathbf{s}_\beta, t_\alpha, t_\beta; \theta) w_{ij, \alpha\beta}, \quad (2.57)$$

où

$$\Lambda = \left\{ (i, j, \alpha, \beta) : \begin{array}{lll} i = j = 1, \dots, p, & \alpha = 1, \dots, n-1, & \beta = \alpha + 1, \dots, n, \\ i = 1, \dots, p-1, & j = i+1, \dots, p, & \alpha, \beta = 1, \dots, n. \end{array} \right.$$

Cette formulation est vraie pour les modèles de fonctions de covariances croisées symétriques puisque dans ce cas $l(i, j, \mathbf{s}_\alpha, \mathbf{s}_\beta, t_\alpha, t_\beta; \theta) = l(j, i, \mathbf{s}_\alpha, \mathbf{s}_\beta, t_\alpha, t_\beta; \theta)$. La log-vraisemblance wpl est donc la somme des contributions des éléments directs ($i = j$) et des éléments croisés ($i \neq j$). Une formulation un peu plus générale est donnée dans ?(papier pas encore soumis de Bevilacqua) pour les modèles non symétriques.

La pondération peut prendre différentes formes comme une fonction de corrélation à support compact ou plus simplement la fonction indicatrice qui donne un poids de 1 à toutes les paires dont la distance spatio-temporelle est inférieure à une borne, et un poids nul aux autres, telle que :

$$w_{ij, \alpha\beta} = \begin{cases} 1 & \text{si } \mathbf{h} = \|\mathbf{s}_\alpha - \mathbf{s}_\beta\| \leq \kappa_{ij}^S \text{ et } u = |t_\alpha - t_\beta| \leq \kappa_{ij}^T, \\ 0 & \text{sinon.} \end{cases} \quad (2.58)$$

Cette borne $\kappa_{ij} = (\kappa_{ij}^S, \kappa_{ij}^T)$ peut être la même pour toutes les variables ou être définie pour chaque cas ij , avec $i, j = 1, \dots, p$. Le choix de cette borne est déjà compliqué dans le cas univarié et il le sera d'avantage dans le contexte multivarié. [Bevilacqua et al. \[2012\]](#) proposent dans le cas univarié un critère basé sur la minimisation de la trace de la matrice de Godambe (généralisation de la matrice de Fisher pour les vraisemblances composites), ce qui revient à minimiser la variance des estimateurs. Malheureusement cette méthode est assez lourde en terme de calculs et délicate à mettre en œuvre pour les covariances de type Matérn lorsque le paramètre de régularité n'est pas fixé. Une autre stratégie, qui s'adapte aussi bien au cas univarié comme multivarié, est de fixer la ou les bornes de façon à couvrir une certaine proportion de la portée pratique de la fonction de covariance ou covariance croisée. La portée pratique est estimée empiriquement par une méthode de moindres carrés.

Enfin, ?(même ref de Bevilacqua pas encore dispo) montrent que l'estimateur du maximum de vraisemblance marginale par paires est consistant et asymptotiquement distribué selon une loi normale (sous l'hypothèse asymptotique d'accroissement du domaine). Dans ce cas, l'inverse de la matrice d'information de Godambe G , définie comme

$$G(\theta) = H(\theta)J(\theta)^{-1}H(\theta)^T, \quad (2.59)$$

où

$$H(\theta) = -\mathbb{E}[\nabla^2 \text{wpl}(\theta)], \quad J(\theta) = \mathbb{E}[\nabla \text{wpl}(\theta) \nabla \text{wpl}(\theta)^T], \quad (2.60)$$

est la variance asymptotique de l'estimateur du maximum de vraisemblance marginale par paires.

2.3 Références

- Abramowitz, M. and Stegun, I. A. (1972). *Handbook of mathematical functions : with formulas, graphs, and mathematical tables*. Number 55. Courier Dover Publications. [27](#)
- Allard, D. (2012). Modeling spatial and spatio-temporal non gaussian processes. In Porcu E., Montero J.M., S. M., editor, *Advances and Challenges in Space-time Modelling of Natural Events*, pages 141–164. Springer. [19](#)
- Apanasovich, T. V., Genton, M. G., and Sun, Y. (2012). A valid matérn class of cross-covariance functions for multivariate random fields with any number of components. *Journal of the American Statistical Association*, 107(497) :180–193. [27](#), [28](#)
- Banerjee, S. and Gelfand, A. (2003). On smoothness properties of spatial processes. *Journal of Multivariate Analysis*, 84(1) :85–100. [27](#)
- Bevilacqua, M., Fassò, A., Gaetan, C., Porcu, E., and Velandia, D. (2015). Covariance tapering for multivariate gaussian random fields estimation. *Statistical Methods & Applications*, pages 1–17. [32](#)
- Bevilacqua, M. and Gaetan, C. (2014). Comparing composite likelihood methods based on pairs for spatial gaussian random fields. *Statistics and Computing*, pages 1–16. [33](#)
- Bevilacqua, M., Gaetan, C., Mateu, J., and Porcu, E. (2012). Estimating space and space-time covariance functions for large data sets : a weighted composite likelihood approach. *Journal of the American Statistical Association*, 107(497) :268–280. [35](#)
- Bochner, S. (1955). Harmonic analysis and the theory of probability. [22](#)
- Chilès, J. and Delfiner, P. (1999). *Geostatistics : Modeling Spatial Uncertainty*. New York : Wiley. [22](#)
- Cramér, H. (1940). On the theory of stationary random processes. *Annals of Mathematics*, pages 215–230. [22](#)
- Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448) :1330–1339. [23](#)
- Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons. [27](#)
- Daley, D. J., Porcu, E., and Bevilacqua, M. (2014). Classes of compactly supported covariance functions for multivariate random fields. *Stochastic Environmental Research and Risk Assessment*, 29(4) :1249–1263. [32](#)

- De Iaco, S., Myers, D., Palma, M., and Posa (2013). Using simultaneous diagonalization to identify a space-time linear coregionalization model. *Mathematical Geosciences*, 45 :69–86. [25](#)
- Feller, V. (1966). *An Introduction to Probability Theory and Its Applications (Vol. II)*. J. Wiley & sons. [29](#)
- Gneiting, T. (1997). Normal scale mixtures and dual probability densities. *Journal of Statistical Computation and Simulation*, 59(4) :375–384. [29](#)
- Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458) :590–600. [23](#), [29](#), [30](#)
- Gneiting, T., Genton, M., and Guttorp, P. (2007). Geostatistical space-time models, stationarity, separability and full symmetry. *Statistical Methods for Spatio-Temporal Systems*, pages 151–175. [25](#)
- Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491). [27](#), [28](#)
- Gneiting, T. and Schlather, M. (2004). Stochastic models that separate fractal dimension and the hurst effect. *SIAM Review*, 96(2) :269–282. [29](#)
- Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484) :1545–1555. [32](#)
- Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields : the stochastic partial differential equation approach. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 73(4) :423–498. [27](#)
- Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80(1) :221–39. [33](#)
- Schlather, M. (2010). Some covariance models based on normal scale mixtures. *Bernoulli*, 16(3) :780–797. [23](#)
- Schoenberg, I. J. (1938). Metric spaces and completely monotone functions. *Annals of Mathematics*, pages 811–841. [29](#)
- Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1) :5–42. [33](#)
- Wackernagel, H. (2003). *Multivariate geostatistics*. Springer. [22](#), [24](#), [32](#), [34](#)

Whittle, P. (1963). Stochastic processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2) :974–994. [26](#)

Yaglom, A. M. (1987). *Correlation theory of stationary and related random functions*, Vol. I : *Basic Results* ; Vol. II : *Supplementary notes and references*. Springer, New York. [22](#)

Chapitre 3

Fully Nonseparable Cross-Covariance Functions for Multivariate Spatio-Temporal Data

Sommaire

3.1 Introduction	40
3.2 A class of non-separable space-time cross-covariance functions	44
3.3 Estimation	47
3.4 Simulation study	49
3.4.1 Setting the optimal window	50
3.4.2 Efficiency of the estimation procedure	51
3.4.3 Assessing the predictive performances	52
3.5 Western France weather dataset	58
3.6 Discussion	64
3.7 Références	67

Ce chapitre est le fruit du travail effectué en collaboration avec Denis Allard et Emilio Porcu. Il a donné lieu à un article soumis à Spatial Statistics, et accepté. La version en ligne est disponible à l'adresse suivante : <http://dx.doi.org/10.1016/j.spasta.2016.02.004>

Abstract

Multivariate space-time data are increasingly available in various scientific disciplines. When analyzing these data, one of the key issues is to describe the multivariate space-time dependencies. Under the Gaussian framework, one needs to propose relevant models for multivariate space-time covariance functions, i.e. matrix-valued mappings with the additional requirement of non-negative definiteness. We propose a flexible parametric class of cross-covariance functions for multivariate space-time Gaussian random fields. Space-time components belong to the (univariate) Gneiting class of space-time covariance functions, with Matérn or Cauchy covariance functions in the spatial margins. The smoothness and scale parameters can be different for each variable. We provide sufficient conditions for positive definiteness. A simulation study shows that the parameters of this model can be efficiently estimated using weighted pairwise likelihood, which belongs to the class of composite likelihood methods. We then illustrate the model on a French dataset of weather variables.

Keywords Composite likelihood ; Matérn covariance ; Multivariate Gaussian processes ; Separability ; Spatio-temporal processes ; Spatio-temporal geostatistics.

3.1 Introduction

Environmental and climate sciences provide an increasing amount of multivariate data indexed by space-time coordinates. For statisticians analyzing these data, one of the key issues is to model the space-time dependence structure, not only within each variable, but also between the variables. This requires models that allow for different range and smoothness parameters for each variable, and whose parameters can be accurately estimated. Let us recall some expository material that will be needed for the presentation of our results.

Consider a p -dimensional multivariate random field $\mathbf{Y}(\mathbf{x}) = \{Y_1(\mathbf{x}), \dots, Y_p(\mathbf{x})\}^\top$, where $Y_i(\mathbf{x})$ represents the i -th variable, $i = 1, \dots, p$, and $\mathbf{x} = (\mathbf{s}, t) \in D \times T \subset \mathbb{R}^{d+1}$, $d \geq 1$, where $\mathbf{s} \in D \subset \mathbb{R}^d$ is a vector of spatial coordinates and $t \in T \subset \mathbb{R}$ is time. Let us further assume that $\mathbf{Y}(\mathbf{x})$ can be decomposed into the sum of a deterministic and a random component,

$$\mathbf{Y}(\mathbf{x}) = \boldsymbol{\mu}(\mathbf{x}) + \mathbf{Z}(\mathbf{x}), \quad \mathbf{x} \in D \times T,$$

where $\boldsymbol{\mu}(\cdot)$ is a trend function and $\mathbf{Z}(\cdot)$ a zero mean multivariate Gaussian stationary process. Under the assumptions of Gaussianity and stationarity, the process $\mathbf{Z}(\cdot)$ is completely

characterized by its matrix-valued covariance function $\mathbf{C}(\mathbf{h}, u) = [C_{ij}(\mathbf{h}, u)]_{i,j=1}^p$, which depends solely on the space-time lag, $\mathbf{k} = (\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R}$:

$$\text{Cov}\{Z_i(\mathbf{s}, t), Z_j(\mathbf{s} + \mathbf{h}, t + u)\} = C_{ij}(\mathbf{h}, u), \quad i, j = 1, \dots, p, \quad \mathbf{s}, \mathbf{s} + \mathbf{h} \in D, \quad t, t + u \in T.$$

Cross-covariance functions are not symmetric, but they are invariant with respect to the joint exchange of the variables and the sign of the lag $\mathbf{k}$: $C_{ij}(\mathbf{k}) = C_{ji}(-\mathbf{k})$, for all $\mathbf{k} \in \mathbb{R}^{d+1}$. Full symmetry is a more restrictive assumption for which the following relationships are also verified : $C_{ij}(\mathbf{k}) = C_{ij}(-\mathbf{k}) = C_{ji}(\mathbf{k})$. For a complete review on multivariate random fields modeling, with a particular focus on spatial cross-covariance functions, we refer the reader to [Genton and Kleiber \[2015\]](#) with the associated discussions.

Our goal is to elaborate valid, flexible parametric classes of space-time matrix-valued covariance functions for $\mathbf{Z}(\mathbf{x})$, i.e. matrix-valued covariance functions that verify the well-known requirement of non-negative definiteness : for any $n \in \mathbb{N}$, for any finite set of points $(\mathbf{s}_1, t_1), \dots, (\mathbf{s}_n, t_n)$ and for any vector $\boldsymbol{\lambda} \in \mathbb{R}^{np}$, we have $\boldsymbol{\lambda}^\top \boldsymbol{\Sigma} \boldsymbol{\lambda} \geq 0$, where $\boldsymbol{\Sigma}$ is a $np \times np$ matrix with $n \times n$ block elements of $p \times p$ matrices $\mathbf{C}(\mathbf{s}_\alpha - \mathbf{s}_\beta, t_\alpha - t_\beta)$, with $\alpha, \beta = 1, \dots, n$.

Following [Genton and Kleiber \[2015\]](#) and [Gelfand and Banerjee \[2010\]](#), a multivariate space-time covariance model is said to be separable when it is obtained through the product of a $p \times p$ covariance matrix $\mathbf{A} = [A_{ij}]_{i,j=1}^p$ and a valid univariate space-time correlation function $\rho_{\text{ST}}(\cdot)$ on $D \times T$,

$$\mathbf{C}(\mathbf{k}) = \mathbf{A} \rho_{\text{ST}}(\mathbf{k}), \quad \mathbf{k} \in D \times T. \quad (3.1)$$

When $\rho_{\text{ST}}(\cdot)$ is also space-time separable, the covariance matrix of $\mathbf{Z}$ reduces to

$$\boldsymbol{\Sigma} = \mathbf{A} \otimes \mathbf{C}_S \otimes \mathbf{C}_T, \quad (3.2)$$

where $\otimes$ is the Kronecker product and $\mathbf{C}_S, \mathbf{C}_T$ are respectively spatial and temporal covariance matrices [[Cressie and Wikle, 2011](#)] associated to univariate spatial and temporal covariance functions. Separability induces reduced number of parameters and faster computation of the inverse and of the determinant of the matrix $\boldsymbol{\Sigma}$. Often, separability is an overly simplified assumption for weather and climate data. Space-time separability is equivalent to conditional independence between $Z(\mathbf{s}, t)$ and $Z(\mathbf{s}', t')$ given $Z(\mathbf{s}', t)$ (or $Z(\mathbf{s}, t')$), since in this case we have

$$\text{Var}\{Z(\mathbf{s}, t)\} \text{Cov}\{Z(\mathbf{s}, t), Z(\mathbf{s}', t')\} = \text{Cov}\{Z(\mathbf{s}, t), Z(\mathbf{s}', t)\} \text{Cov}\{Z(\mathbf{s}', t), Z(\mathbf{s}', t')\},$$

$(\mathbf{s}, t), (\mathbf{s}', t') \in D \times T$. As a consequence, there is a proportional relationship between $C(\mathbf{h}, u)$ and $C(\mathbf{h}', u)$ for two fixed spatial lags $\mathbf{h}, \mathbf{h}' \in \mathbb{R}^d, u \in \mathbb{R}$, which implies that separable covariances cannot capture sophisticated interactions between space and time. Moreover, in a multivariate framework, separability between variables and space-time variations implies that all variables are characterized by the same space-time correlation function. An

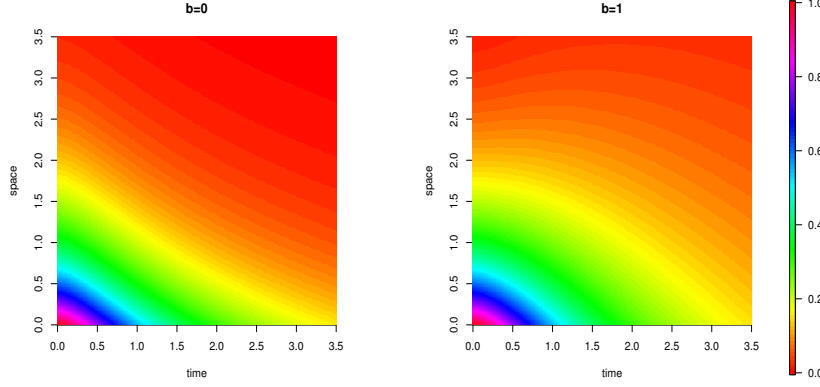


FIGURE 3.1 – An example of a Gneiting space-time covariance function $\mathcal{G}(\mathbf{h}, u) = (0.8|u|^{1.5} + 1)^{-1} \exp\{-\|\mathbf{h}\|/(0.8|u|^{1.5} + 1)^{-b/2}\}$. The case $b = 0$ corresponds to separability.

approach based on the asymptotic distribution of the sample cross-covariance estimator to test for separability and for full symmetry is proposed in [Li et al. \[2008\]](#). This allows the practitioner to select among the important dependence structures and to make the appropriate modeling choice. Often, separability must be rejected and models that do not separate space, time and variable index must be defined.

[Gneiting \[2002\]](#) proposed a class of univariate fully symmetric space-time covariances, that has now become the standard class of models for univariate space-time Gaussian random fields in geostatistical applications; see also [Gneiting et al. \[2007\]](#) and references therein. The Gneiting class of space-time covariances is defined as

$$\mathcal{G}(\mathbf{h}, u) = \frac{\sigma^2}{\psi(u^2)^{d/2}} \varphi\left(\frac{\|\mathbf{h}\|^2}{\psi(u^2)}\right), \quad (\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R}, \quad (3.3)$$

where $\varphi : [0, \infty) \rightarrow \mathbb{R}$ is completely monotone on the positive real line, with $\varphi(0) < \infty$, ψ is a positive function whose derivative is completely monotone on the positive real line and σ^2 is a variance parameter. See [Porcu and Zastavnyi \[2011\]](#) for more relaxed necessary conditions and the complete characterization of this class. Figure 3.1 depicts an example of a Gneiting space-time covariance function.

In spatial context, multivariate models for which space and variables are not separable are easily built by combining separable ones. In the linear model of coregionalization (LMC, [Goulard and Voltz \[1992\]](#); [Wackernagel \[2003\]](#)), the matrix-valued covariance function is defined as a linear combination of separable models : $\mathbf{C}(\mathbf{h}) = \sum_{k=1}^K \mathbf{B}_k \rho_k(\mathbf{h})$, where $\rho_k(\mathbf{h})$ are spatial covariance functions with $\rho_k(\mathbf{0}) = 1$ and $\mathbf{B}_k$ are positive definite matrices. Drawbacks of this construction have been discussed in [Gneiting et al. \[2010\]](#) and in [Daley et al. \[2014\]](#). [De Iaco et al. \[2013\]](#) proposed a space-time extension of this construction. A major drawback of this approach is that the smoothness of any component of the multivariate random field is that of the roughest process associated to $\rho_k(\cdot)$, $k = 1, \dots, K$. [Apanasovich et al. \[2012\]](#) introduced a valid parametric family of cross-covariance functions

for multivariate spatial random fields where each component has a covariance function from a Matérn class.

In space-time context, [Apanasovich and Genton \[2010\]](#) proposed valid non-separable matrix-valued space-time covariance functions through additional latent dimensions, one for each of the p variables, that represent the variables to be modeled. These dimensions are then used in a similar way as that of time within the Gneiting class of covariance functions. Climate and weather variables need flexible space-time models able to capture different regularity properties from one variable to another. Under the framework proposed in [Apanasovich and Genton \[2010\]](#), this can be obtained through a LMC-type construction, at the cost of a significant increase of the number of parameters. In order to keep a reasonably low number of parameters, we argue in favor of a different approach.

The Matérn family [[Matérn, 1986](#)] of covariance functions has become extremely popular in geostatistical applications. Its expression is

$$\mathcal{M}(\mathbf{h}; r, \nu) = \frac{\sigma^2 2^{1-\nu}}{\Gamma(\nu)} (r \|\mathbf{h}\|)^\nu \mathcal{K}_\nu(r \|\mathbf{h}\|), \quad \mathbf{h} \in \mathbb{R}^d,$$

where $r > 0$ is a scale parameter ($1/r$ is often called the range) and $\mathcal{K}_\nu$ denotes the modified Bessel function of the second kind of order ν [[DLMF, 2015](#)]. The smoothness parameter $\nu > 0$ is directly related to the regularity of the underlying random field. Specifically, a random field $Y(\mathbf{x})$ with Matérn covariance is m times mean square differentiable if and only if $\nu \geq m$. As $\nu \rightarrow \infty$, the random field is infinitely mean square differentiable and its associated covariance function tends to the so-called Gaussian covariance function $C(\mathbf{h}) = \exp(-r^2 \|\mathbf{h}\|^2)$, $\mathbf{h} \in \mathbb{R}$. Small values of ν yield rougher random fields; in particular $\nu = 1/2$ corresponds to the exponential covariance function $C(\mathbf{h}) = \exp(-r \|\mathbf{h}\|)$ that is mean square continuous but not mean square differentiable at the origin. When $\nu = k + 1/2$ and k is a positive integer, the Matérn model reduces to the product of the negative exponential with a polynomial of degree k .

A multivariate version of the Matérn family has been proposed in [Gneiting et al. \[2010\]](#) and further extended in [Apanasovich et al. \[2012\]](#). These constructions have the same general structure with

$$C_{ij}(\mathbf{h}) = \sigma_i \sigma_j \rho_{ij} \mathcal{M}(\mathbf{h}; r_{ij}, \nu_{ij}), \quad \mathbf{h} \in \mathbb{R}^d, \quad (3.4)$$

with different restrictions on the parameters $\{r_{ij}, \nu_{ij}, \rho_{ij}\}_{i,j=1,\dots,p}$ to ensure the validity of the matrix-valued covariance function $\mathbf{C}(\mathbf{h}) = [C_{ij}(\mathbf{h})]_{i,j=1}^p$. In this model, each marginal component of $\mathbf{Z}$, i.e. each covariance function $C_{ii}(\mathbf{h})$, has a different smoothness parameter, while allowing some cross correlations between the variables. Sufficient conditions have been obtained using scalar mixtures [[Gneiting et al., 2010](#); [Porcu and Zastavnyi, 2011](#); [Schlather, 2010](#)] while necessary and sufficient conditions are provided in [Gneiting et al. \[2010\]](#) for $p = 2$ only.

As an alternative to the Matérn class, it is also possible to build a multivariate version of the generalized Cauchy family, having expression

$$\mathcal{C}(\mathbf{h}; r, \beta, \lambda) = \left(1 + r \|\mathbf{h}\|^\lambda\right)^{-\beta/\lambda}, \quad \mathbf{h} \in \mathbb{R}^d,$$

with $0 < \lambda \leq 2$ and $\beta, r > 0$. In [Gneiting and Schlather \[2004\]](#) it is shown that λ characterizes the roughness of the associated random field. Higher values of λ yielding smoother fields, while β parametrizes the dependence at large distances. Another parameterization, which will be used in the rest of this work, is obtained by replacing β/λ by $\nu > 0$. The isotropic Matérn and Cauchy covariance functions are related to the completely monotone functions of Table 1 in [Gneiting \[2002\]](#) by the equation $C(\mathbf{h}) = \varphi(\|\mathbf{h}\|^2)$ [[Schoenberg, 1938](#)].

In this article we propose a class of non-separable multivariate space-time covariance models for any number of variables. The general structure of this class is $\mathbf{C}(\mathbf{h}, u) = [C_{ij}(\mathbf{h}, u)]_{i,j=1}^p$ with

$$C_{ij}(\mathbf{h}, u) = \frac{\sigma_i \sigma_j}{\psi(u^2)^{d/2}} \rho_{ij} \varphi_{ij} \left(\frac{\|\mathbf{h}\|^2}{\psi(u^2)} \right), \quad 1 \leq i, j \leq p, \quad (\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R}.$$

It is an extension of the Gneiting class $\mathcal{G}$ to the multivariate case, where the completely monotone functions φ_{ij} have different parameters for each variable.

In particular, we offer sufficient conditions for two cases : $\varphi_{ij}(t) = \mathcal{M}(t^2; r_{ij}, \nu_{ij})$ and $\varphi_{ij}(t) = \mathcal{C}(t^2; r_{ij}, \nu_{ij}, \lambda)$. The parameterization is inspired by the multivariate Matérn structure in [Gneiting et al. \[2010\]](#) and [Apanasovich et al. \[2012\]](#) for spatial Gaussian random fields. For both classes, each variable has its own range and own degree of smoothness, while allowing for cross-correlation. The remainder of the article is organized as follows. Section 2 presents the new multivariate space-time classes. Sufficient conditions that provide valid matrix-valued covariance functions are established and proved in the Appendix. An estimation technique, based on pairwise composite likelihood, is then presented in Section 3. Section 4 is devoted to a simulation study in order to validate the estimation procedure. In Section 5, the Matérn model is applied to a multivariate space-time dataset of weather variables in France. It is shown that the proposed model offers substantial improvements to separable models and also to more parsimonious models with equal range and smoothness parameters for all variables.

3.2 A class of non-separable space-time cross-covariance functions

The mixture approach described in [Gneiting et al. \[2010\]](#) to build Matérn matrix-valued covariance functions for multivariate random fields is based on the stability properties of positive definite functions [[Reisert and Burkhardt, 2007](#), Proposition 3.1]. It uses the scale mixture representation of the Matérn covariance function [[Gradshteyn and Ryzhik, 2007](#),

Eq. 3.471.9]. We extend it to multivariate space-time random fields to build valid matrix-valued covariance functions in $\mathbb{R}^d \times \mathbb{R}$. It can also be adapted to other covariance functions allowing mixture representations, such as the Cauchy model. Let us first recall the mixture representation in the space-time framework.

Lemma 1 *Let the matrix-valued function $\mathbf{m}(\xi) : \mathcal{Q} \subset \mathbb{R} \rightarrow \mathbb{R}^p \times \mathbb{R}^p$ be symmetric and non-negative definite for all $\xi \in \mathcal{Q}$. Let $C_\xi : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}$ be a univariate space-time covariance function for any fixed $\xi \in \mathcal{Q}$. Suppose furthermore that for all $i, j = 1, \dots, p$ and fixed $(\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R}$ the product $m_{ij}(\cdot)C_\xi(\cdot)$ is integrable with respect to a positive measure F on $\mathcal{Q}$. Then, the matrix-valued function $\mathbf{C}(\mathbf{h}, u) = [C_{ij}(\mathbf{h}, u)]_{i,j=1}^p$ defined through*

$$C_{ij}(\mathbf{h}, u) = \int_{\mathcal{Q}} m_{ij}(\xi) C_\xi(\mathbf{h}, u) F(d\xi), \quad 1 \leq i, j \leq p, \quad (3.5)$$

is a valid p -variate matrix-valued covariance function on $\mathbb{R}^d \times \mathbb{R}$.

This lemma is a direct application of Theorem 1 in [Porcu and Zastavnyi \[2011\]](#). With appropriate choices for $C_\xi(\cdot)$ and $\mathbf{m}(\cdot)$ the following results can be established.

Theorem 1 *Let $\psi(t), t \geq 0$, be a positive function with a completely monotone derivative. The multivariate Gneiting-Matérn space-time model denoted $\mathbf{C}^{\mathcal{M}} = [C_{ij}^{\mathcal{M}}(\cdot, \cdot)]_{i,j=1}^p$, with*

$$C_{ij}^{\mathcal{M}}(\mathbf{h}, u) = \frac{\sigma_i \sigma_j}{\psi(u^2)^{d/2}} \rho_{ij} \mathcal{M} \left(\frac{\mathbf{h}}{\psi(u^2)^{1/2}}; r_{ij}, \nu_{ij} \right), \quad (\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R}, \quad (3.6)$$

defines a valid matrix-valued covariance function if, for all $i, j = 1, \dots, p$,

$$\begin{aligned} r_{ij} &= \{(r_i^2 + r_j^2)/2\}^{1/2}, \\ \nu_{ij} &= (\nu_i + \nu_j)/2, \\ \rho_{ij} &= \beta_{ij} \frac{\Gamma(\nu_{ij})}{\Gamma(\nu_i)^{1/2} \Gamma(\nu_j)^{1/2}} \frac{r_i^{\nu_i} r_j^{\nu_j}}{r_{ij}^{2\nu_{ij}}}, \end{aligned}$$

with $r_i, \nu_i > 0$ for all $i = 1, \dots, p$, and where $\boldsymbol{\beta} = [\beta_{ij}]_{i,j=1}^p$ is a correlation matrix.

Theorem 2 *Let $\psi(t), t \geq 0$, be a positive function with a completely monotone derivative. The multivariate Gneiting-Cauchy space-time model denoted $\mathbf{C}^{\mathcal{C}} = [C_{ij}^{\mathcal{C}}(\cdot, \cdot)]_{i,j=1}^p$, with*

$$C_{ij}^{\mathcal{C}}(\mathbf{h}, u) = \frac{\sigma_i \sigma_j}{\psi(u^2)^{d/2}} \rho_{ij} \mathcal{C} \left(\frac{\mathbf{h}}{\psi(u^2)^{1/2}}; r_{ij}, \nu_{ij}, \lambda \right), \quad (\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R}, \quad (3.7)$$

is a valid matrix-valued covariance function if, for all $i, j = 1, \dots, p$,

$$\begin{aligned} r_{ij} &= \{(r_i^{-1} + r_j^{-1})/2\}^{-1}, \\ \nu_{ij} &= (\nu_i + \nu_j)/2, \\ \rho_{ij} &= \beta_{ij} \frac{\Gamma(\nu_{ij})}{\Gamma(\nu_i)^{1/2} \Gamma(\nu_j)^{1/2}} \frac{r_{ij}^{\nu_{ij}}}{(r_i^{\nu_i} r_j^{\nu_j})^{1/2}}, \end{aligned}$$

with $r_i, \nu_i > 0$ for all $i = 1, \dots, p$ and $0 < \lambda \leq 2$, and where $\boldsymbol{\beta} = [\beta_{ij}]_{i,j=1}^p$ is a correlation matrix.

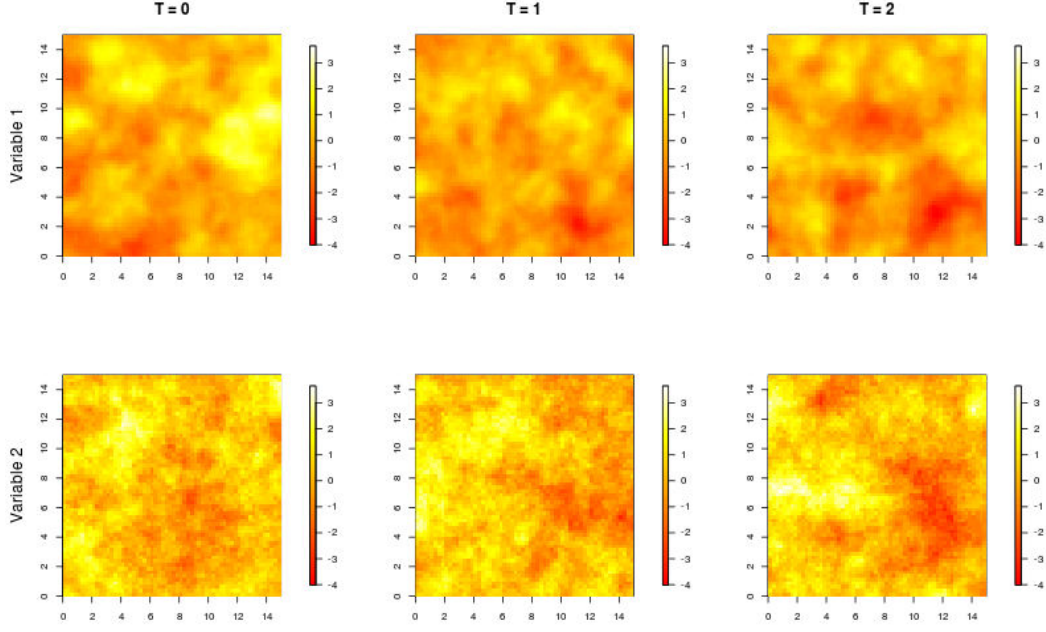


FIGURE 3.2 – Bivariate Gneiting-Matérn space-time Gaussian random field at 3 consecutive instants. Top row : smoother component ($\nu_1 = 1.5$ and $r_1 = 1$); bottom row : rougher component ($\nu_2 = 0.5$ and $r_2 = 0.5$). The co-located correlation parameter ρ_{12} is set to 0.5.

The proofs are deferred to the Appendix. Each marginal space-time covariance function $C_{ii}^{\mathcal{F}}(\cdot, \cdot)$, $i = 1, \dots, p$, $\mathcal{F} \in \{\mathcal{M}, \mathcal{C}\}$, belongs to the Gneiting class with different regularity parameters ν_i , $i = 1, \dots, p$ for its associated Matérn, respectively Cauchy, spatial covariance function. The Gneiting-Matérn model is a space-time generalization of one particular case of the multivariate spatial “flexible model” presented in [Apanasovich et al. \[2012\]](#). Figure 3.2 shows a realization of a bivariate Gaussian field with Gneiting-Matérn space-time covariance at instants $t \in \{0, 1, 2\}$. In this example, the first component is smooth, with $\nu_1 = 1.5$ and scale parameter $r_1 = 1$, whereas the second component is rough, with $\nu_2 = 0.5$ and $r_2 = 0.5$. The co-located correlation parameter is $\rho_{12} = 0.5$.

Positive functions with completely monotone derivative, sometimes referred to as Bernstein functions [[Porcu and Schilling, 2011](#)], are related to isotropic variograms by the following relationship : if a function $\psi(x)$, $x \geq 0$, belongs to the class of Bernstein functions, the function $\gamma(\mathbf{h}) = \psi(\|\mathbf{h}\|^2) - c$, $\mathbf{h} \in \mathbb{R}^d$ is a valid isotropic variogram for some constant c . For the rest of this work, we will consider the parametric form advocated in [Gneiting \[2002\]](#) and [Schlather \[2010\]](#)

$$\psi(x) = (\alpha x^a + 1)^b, \quad x \geq 0,$$

with $\alpha > 0, 0 < a \leq 1, 0 \leq b \leq 1$. With this choice (and a slight abuse of notation, since the Matérn covariance does not have a parameter λ), the above models become

$$C_{ij}(\mathbf{h}, u) = \frac{\sigma_i \sigma_j}{(\alpha |u|^{2a} + 1)^{bd/2}} \rho_{ij} \mathcal{F} \left(\frac{\mathbf{h}}{(\alpha |u|^{2a} + 1)^{b/2}}; r_{ij}, \nu_{ij}, \lambda \right), \quad (\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R}, \quad (3.8)$$

with $\mathcal{F} \in \{\mathcal{M}, \mathcal{C}\}$ and the parameter restrictions stated in Theorems 1 and 2. Note that in the parameterization (3.8), the space-time non-separability parameter b acts both on the spatial covariance $\mathcal{F}$ and on the temporal one. A major drawback of this parameterization is that in case of space-time separability, i.e. when $b = 0$, the temporal covariance is equal to 1 for all $u \in \mathbb{R}$. Following Gneiting [2002] a reparameterization is useful. Multiplying (3.8) by the temporal covariance function $(\alpha|u|^{2a} + 1)^{-\delta/2}$, $u \in \mathbb{R}$, with $\delta > 0$ and replacing the exponent $\delta + bd/2$ by τ leads to the parametric family, which will be used in the rest of this work

$$C_{ij}(\mathbf{h}, u) = \frac{\sigma_i \sigma_j}{(\alpha|u|^{2a} + 1)^\tau} \rho_{ij\mathcal{F}}\left(\frac{\mathbf{h}}{(\alpha|u|^{2a} + 1)^{b/2}}; r_{ij}, v_{ij}, \lambda\right), \quad (\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R}, \quad (3.9)$$

with $\tau \geq bd/2$. This parameterization provides independent and interpretable parameters for the spatial scale and smoothness governed respectively by the parameters r_i and v_i , the temporal scale α , smoothness a and the separability b . Note that the above class of models is fully symmetric, i.e. $C_{ij}(\mathbf{h}, u) = C_{ji}(\mathbf{h}, u) = C_{ij}(\mathbf{h}, -u) = C_{ij}(-\mathbf{h}, u)$ for all $i, j = 1, \dots, p$.

3.3 Estimation

For a Gaussian random field, maximum full likelihood (FL) requires the evaluation of the determinant and the inverse of the $np \times np$ covariance matrix, where n is the number of space-time points and p the number of variables. The computational cost being of the order of $\mathcal{O}((np)^3)$ for both operations, a maximum full likelihood is unfeasible for large datasets, and even for datasets of moderate size as soon as the number of variables increases.

Composite likelihood (CL) methods [Lindsay, 1988] have been proposed to perform efficient estimation with less time-consuming computational steps, and to guarantee good asymptotic properties. Composite likelihoods are products of smaller likelihoods defined on certain subsets of data such as marginal or conditional events, which are easy to compute. See Varin et al. [2011] for an overview. Following Bevilacqua and Gaetan [2015], we choose the pairwise marginal Gaussian likelihoods computed on all pairs of data $\{Z_i(\mathbf{s}_\alpha, t_\alpha), Z_j(\mathbf{s}_\beta, t_\beta)\}$, where $i, j = 1, \dots, p$ and $\alpha, \beta = 1, \dots, n$, n being the number of sites. The negative log-likelihood of such pairs is

$$l(i, j, \mathbf{s}_\alpha, \mathbf{s}_\beta, t_\alpha, t_\beta; \theta) = \frac{1}{2} \left\{ \log \Delta_{ij, \alpha\beta} + \frac{A_{ij, \alpha\beta}}{\Delta_{ij, \alpha\beta}} \right\}$$

where $\Delta_{ij, \alpha\beta} = \sigma_i^2 \sigma_j^2 - C_{ij}(\mathbf{h}, u)^2$ and $A_{ij, \alpha\beta} = \sigma_j^2 Z_i(\mathbf{s}_\alpha, t_\alpha)^2 - 2C_{ij}(\mathbf{h}, u) Z_i(\mathbf{s}_\alpha, t_\alpha) Z_j(\mathbf{s}_\beta, t_\beta) + \sigma_i^2 Z_j(\mathbf{s}_\beta, t_\beta)^2$, with $\mathbf{h} = \|\mathbf{s}_\alpha - \mathbf{s}_\beta\|$, $u = |t_\alpha - t_\beta|$. In the special case of multivariate symmetric models, the Weighted Pairwise log-Likelihood (WPL) is thus

$$\text{wpl}(\theta) = \sum_{(i, j, \alpha, \beta) \in \Lambda} l(i, j, \mathbf{s}_\alpha, \mathbf{s}_\beta, t_\alpha, t_\beta; \theta) w_{\alpha\beta}, \quad \theta \in \Theta, \quad (3.10)$$

where

$$\begin{aligned} \Lambda &= \{i = j = 1, \dots, p, \quad \alpha = 1, \dots, n-1, \quad \beta = \alpha + 1, \dots, n\} \\ &\cup \{i = 1, \dots, p-1, \quad j = i + 1, \dots, p, \quad \alpha, \beta = 1, \dots, n\}, \end{aligned}$$

and $\Theta \subset \mathbb{R}^q$ is the space of parameters for which the model in (3.9) is valid, with $q = (p+2)(p+3)/2$ for the Gneiting-Matérn class and $q = (p+2)(p+3)/2+1$ for the Gneiting-Cauchy class.

The computational cost of WPL is of the order of $\mathcal{O}((np)^2)$ when considering all possible pairs. It can be significantly reduced with an adequate choice of the weights. In this paper, we have chosen cut-off weights, that is $w_{\alpha\beta} = 1$ if the space-time lag $(\|\mathbf{h}\|, u) \leq (d_S, d_T)$, where the order is defined pointwise. Otherwise we set $w_{\alpha\beta} = 0$. We use the expression *weighted pairwise likelihood* but it is sometimes called *truncated composite likelihood* or *tapered composite likelihood*. In spite of its simplicity, this weighting scheme provides a significant gain in computational time, and preserves a reasonable level of statistical efficiency, since pairs of observations whose distance is beyond the correlation range are uninformative for the smoothness and range parameters of the covariance function. The choice of $\mathbf{d} = (d_S, d_T)$ will be discussed in details in the next section. At this stage, we can notice that $\mathbf{d}$ is the same for all variables.

In the univariate setting, Bevilacqua et al. [2012] proposed to seek the “optimal” window $\mathbf{d}^*$ such that

$$\mathbf{d}^* = \arg \min_{\mathbf{d}} \text{tr} \left(\mathbf{G}_{np}^{-1}(\mathbf{d}; \theta) \right), \quad \theta \in \Theta, \quad (3.11)$$

where θ is replaced by a consistent estimator of θ , such as for example the weighted least squares estimate based on empirical estimation of the variogram [Cressie, 1993, p. 91]. The value $\mathbf{d}^*$ minimizes the sum of the diagonal elements of the inverse of $\mathbf{G}_{np}^{-1}(\mathbf{d}; \theta)$, which is the asymptotic variance-covariance matrix of the WPL estimators. The matrix $\mathbf{G}_{np}(\mathbf{d}; \theta)$ is the Godambe information matrix, also referred to as the “sandwich” information matrix [Bevilacqua et al., 2012]. The Godambe information matrix of a full log-likelihood function is equal to the Fisher information matrix.

Estimating the parameters from model (3.9) requires to maximize $\text{wpl}(\theta)$ in Θ , when $\text{wpl}(\theta)$ is computed with the appropriate window $\mathbf{d}^*$. Closed form expressions for the theoretical expression of the Godambe matrix are not easily feasible since first and second order derivatives of (4.18) with respect to all parameters are required. We thus favor an empirical approach, based on a simulation study detailed in the next Section. For a given value of $\mathbf{d}^*$, maximizing $\text{wpl}(\theta)$ in Θ is by no means trivial, since model in (3.9) has $(p+2)(p+3)/2$ parameters. Employing blindly an optimization function is deemed to fail, even for the simplest case with $p = 2$.

As a way to alleviate this problem, $\text{wpl}(\theta)$ is maximized sequentially in subspaces of Θ , corresponding to blocks of related parameters, while keeping all other parameters fixed

to the previously attained values. Among the schemes that have been tested, the following combination seemed to lead to the best estimates : for a fixed value of b , first seek the maximum corresponding to the $p(p-1)/2$ correlation parameters β_{ij} ; then, for each variable i , find the maximum for σ_i , followed by (v_i, r_i) ; finally, maximize with respect to the two temporal parameters a and α . One iteration is achieved when all parameters have been estimated once. The likelihood is iteratively maximized until a stopping criterion is reached. We have chosen to stop the maximization whenever the log-likelihood is increased by less than one unit. Other orders have been tested, leading to very close estimates.

The separability parameter b , crucial in this study, is difficult to optimize in the above framework. We often observed unstable maximization and/or estimates reaching the boundary of the parameter space Θ . Since b belongs to the interval $[0, 1]$, we decided to estimate it by a simple grid search. The parameter b is successively fixed to the values $k/10$, with $k = 0, \dots, 10$. For each of these values, the above maximization procedure is performed. Finally $\hat{b}$ is that value of b corresponding to the highest maximized likelihood.

To perform the maximization within the subspaces of Θ , we used the function *optim* with quasi Newton method "BFGS" (published simultaneously in 1970 by Broyden, Fletcher, Goldfarb and Shanno) as implemented in R. Relevant initial values were set by estimating parameters of the corresponding marginal spatial and temporal covariance functions, independently for each variable. Initial values of the correlation coefficients were set to their empirical values.

3.4 Simulation study

We propose a simulation study with three main goals : setting the optimal window $\mathbf{d}^*$, assessing the performance of the estimation procedure detailed in Section 3.3, and measuring the gain in prediction implied by model (3.9), as compared to non-separable or less flexible models. Because of the wide popularity of the Matérn covariance functions, and also for space considerations, we only report results obtained with the multivariate Gneiting-Matérn class (3.6) in Theorem 1. The setting of the simulation study mimics the conditions found for the analysis of weather data as detailed in Section 3.5. Three variables are available during one month (30 days) at 13 locations (the precise locations are shown later, in Figure 3.4). Among these, 11 sites are used for the first two tasks (circles) while two sites are kept for validation (stars). There are 10 years of data, assumed independent. We thus simulate a sample of size 10 from a Gaussian vector with $30 \times 13 \times 3 = 1170$ components. The elements of the 1170×1170 matrix are given by the multivariate space-time Matérn covariance function with unit variance for all variables, i.e.

$$C_{ij}(\mathbf{h}, u) = \frac{\rho_{ij}}{\alpha|u|^{2a+1}} \mathcal{M}\left(\frac{\mathbf{h}}{(\alpha|u|^{2a+1})^{b/2}}; r_{ij}, v_{ij}\right) \quad i, j = 1, \dots, p. \quad (3.12)$$

$\mathbf{d}$	$d_T = 2$ days		$d_T = 5$ days		$d_T = 10$ days	
$d_S = 250$ km	0.120	(7.2%)	0.148	(15.1%)	0.185	(26.2%)
$d_S = 500$ km	0.127	(13.5%)	0.162	(28.3%)	0.203	(49.1%)
$d_S = 750$ km	0.146	(15.4%)	0.196	(33.3%)	0.227	(57.7%)

TABLE 3.1 – Average of the sum of the estimated variances of all parameters for several windows $\mathbf{d} = (d_S, d_T)$. 100 synthetic datasets simulated according to model (3.12). The number in brackets corresponds to the percentage of pairs used in the computation of wpl.

Exact simulations with Cholesky decomposition are feasible, and a total of 100 repetitions are simulated. The reasonable dimension of the matrices (1170×1170) allows to compare WPL inference with a usual likelihood approach, referred to as the Full Likelihood (FL) approach. Parameters have been chosen to match those estimated by an exploratory analysis of the spatial and temporal margins of the dataset (see Table 3.3). In order to account for lack of space-time separability, the parameter b was set equal to 0.8. As already pointed out earlier, we only report results obtained with the Matérn model.

3.4.1 Setting the optimal window

The optimal window $\mathbf{d}^*$ given in Eq. (3.11) minimizes the sum of the estimation variance of every parameter. For several choices of the spatial distance $d_S \in \{250, 500, 750\}$ (expressed in kms), and temporal distance $d_T \in \{2, 5, 10\}$ (in days), the estimation of the parameters is performed on each simulated repetition. Then, empirical estimates of the estimation variances are computed. Their sum is an estimate of the criterion in Eq. (3.11), as recommended in Bevilacqua et al. [2012]. Results are reported in Table 3.1. As the spatial distance increases, the quantity $\mathbf{G}_{np}^{-1}(\mathbf{d}; \hat{\theta})$ increases for all temporal distances. For all spatial distances, the criterion is minimal for the smallest temporal window. Parameters are best estimated if pairs corresponding to moderate distances and to successive days only are used, which corresponds to 7.2% and 13.5% of the number of pairs for the two lowest values of the criterion. Overall, the window $\mathbf{d} = (250\text{km}, 2\text{days})$ is optimal according to this criterion. The criterion computed with $\mathbf{d} = (500\text{km}, 2\text{days})$ is very close to the optimal. The estimated mean and variance of the separability parameter b for the different space-time windows, is reported in Table 3.2. The picture is now slightly different. Bias and estimation variance are minimal for $\mathbf{d} = (500\text{km}, 2\text{days})$. Considering the importance of the separability parameter, and since the trace criterion for this window is very close to the optimum, the window is set to $\mathbf{d} = (500\text{km}, 2\text{days})$ for the rest of this work. These results are consistent with those reported in Bevilacqua et al. [2012], where it is observed that the relative efficiency of WPL as a function of the distance first increases to a maximum and then decreases as more distant pairs are added in the WPL.

d	$d_T = 2$ days		$d_T = 5$ days		$d_T = 10$ days	
$d_S = 250$ km	0.694	(0.234)	0.681	(0.268)	0.630	(0.305)
$d_S = 500$ km	0.768	(0.230)	0.731	(0.251)	0.717	(0.259)
$d_S = 750$ km	0.744	(0.245)	0.723	(0.257)	0.696	(0.281)

TABLE 3.2 – Estimated mean and standard deviation (in brackets) of the separability parameter b for several windows $\mathbf{d} = (d_S, d_T)$. Same simulations as in Table 3.1.

3.4.2 Efficiency of the estimation procedure

We now assess the ability of the estimation procedure to estimate efficiently the different parameters, the window being set to $\mathbf{d} = (500\text{km}, 2\text{days})$. This simulation study also allows us to compare the efficiency of WPL with respect to FL for the multivariate Gneiting-Matérn class. We maximize WPL and FL, using the same maximization scheme, on the same set of 100 simulations. As can be observed from Table 3.3, the overall performances are good for both approaches. For most parameters, the difference between the median and the mean is negligible, and differences between the mean and the true value is relatively small for most parameters. The range and smoothness parameters are a notable exception. Their estimators are less biased and more dispersed when using WPL than when using FL. For all variables $i \in \{1, 2, 3\}$, the smoothness parameters, ν_i , tend to be overestimated, while the scale parameters, $1/r_i$, tend to be underestimated. In Zhang [2004], it is shown that these parameters compensate each other and that the simultaneous estimation of both is difficult. Results shown in Table 3.3 confirm our assertions. In accordance with the parameterization chosen in Theorems 1 and 2, the correlation coefficients β_{ij} , with $1 \leq i < j \leq 3$ must belong to the interval $(-1, 1)$ with $\beta_{ii} = 1$, for $i = 1, \dots, 3$, and the matrix $[\beta_{ij}]_{i,j=1}^3$ must be positive definite. The estimators of the correlations coefficients β_{ij} , with $1 \leq i < j \leq 3$, are unbiased for both approaches, but their dispersion is much larger when using WPL than when using FL. We return to this point later. Remember that the separability parameter b is not estimated continuously and that the window $\mathbf{d}$ has been chosen to minimize the bias and the variance of its estimator. Interestingly, it is better estimated using WPL than using FL. With WPL, the median is equal to the true value and it is close to the mean. The separability parameter is estimated to be non null, for all simulated datasets but one.

To complete this comparison, we compute the relative efficiency of WPL with respect to FL. For each approach, and for each of the 15 parameters θ_i , with $i \leq i \leq 15$, the root mean square error, $\text{rmse}_i = \sqrt{\text{bs}_i^2 + \text{sd}_i^2}$, is computed, where the bias is $\text{bs}_i = \tilde{\theta}_i - \theta_i$, and the variance is $\text{sd}_i^2 = \sum_{j=1}^{100} (\hat{\theta}_{j,i} - \tilde{\theta}_i)^2 / 100$, with $\tilde{\theta}_i = \sum_{j=1}^{100} \hat{\theta}_{j,i} / 100$. For each variable, we denote $\text{rmse}_i^{\text{FL}}$ and $\text{rmse}_i^{\text{WPL}}$ the root mean square error corresponding respectively to FL and to

	True	Weighted Pairwise Likelihood						Full Likelihood					
		Min	Q ₁	Median	Mean	Q ₃	Max	Min	Q ₁	Median	Mean	Q ₃	Max
σ_1	1.00	0.89	0.97	1.00	1.00	1.02	1.16	0.93	0.97	0.99	0.99	1.01	1.06
σ_2	1.00	0.91	0.97	1.00	1.00	1.02	1.11	0.95	0.98	0.99	1.00	1.01	1.06
σ_3	1.00	0.92	0.97	0.99	0.99	1.02	1.10	0.94	0.97	0.99	0.99	1.00	1.12
β_{12}	-0.40	-0.61	-0.46	-0.40	-0.40	-0.34	-0.15	-0.45	-0.42	-0.41	-0.41	-0.40	-0.37
β_{13}	-0.40	-0.62	-0.47	-0.41	-0.41	-0.34	-0.21	-0.45	-0.42	-0.41	-0.41	-0.40	-0.36
β_{23}	0.25	-0.05	0.19	0.27	0.26	0.33	0.47	0.21	0.25	0.26	0.26	0.27	0.31
v_1	0.70	0.55	0.69	0.74	0.75	0.82	1.01	0.58	0.72	0.75	0.76	0.79	0.99
v_2	0.80	0.55	0.76	0.83	0.84	0.91	1.17	0.68	0.82	0.87	0.88	0.95	1.17
v_3	0.40	0.28	0.39	0.43	0.42	0.46	0.72	0.30	0.43	0.45	0.45	0.47	0.59
$1/r_1$	250	135	205	229	235	262	388	147	199	216	220	238	333
$1/r_2$	200	128	170	190	192	215	299	131	158	177	178	193	261
$1/r_3$	350	149	284	310	327	356	1017	195	247	277	283	293	1017
α	0.90	0.65	0.84	0.90	0.91	0.99	1.24	0.78	0.91	0.95	0.95	1.00	1.12
a	0.50	0.35	0.46	0.49	0.50	0.54	0.71	0.46	0.49	0.51	0.51	0.53	0.58
b	0.80	0.00	0.60	0.80	0.77	1.00	1.00	0.30	0.50	0.60	0.64	0.70	1.00

TABLE 3.3 – Summary statistics of the estimated parameters for $\mathbf{d} = (500 \text{ km}, 2 \text{ days})$. Same simulations as in Table 3.1.

WPL. The root relative efficiency, defined as

$$\text{rre}_i = \text{rmse}_i^{\text{FL}} / \text{rmse}_i^{\text{WPL}},$$

is reported for each variable in Table 3.4. When $\text{rre} < 1$, FL is more efficient than WPL, and conversely when $\text{rre} > 1$. For all parameters, it was found that FL leads to higher bias than WPL, but that the lower variance associated to FL usually more than compensates this effect. On these simulations, WPL is found to be rather inefficient for the estimation of the correlation parameters β_{ij} , for which the variance associated to WPL is about 30 times that associated to FL. For all other parameters and in particular for regularity and range parameters, WPL provides estimates with a mild loss in efficiency as compared to a full likelihood approach, with a significant gain in terms of computation. Similar results were obtained for the inference of the parameters of Max-Stable processes in [Castruccio et al. \[2015\]](#). Rarely, the relative efficiency is even slightly larger than 1 but this could be due to the moderate number of replicates. In these cases, the lower variance for FL does not compensate the larger bias.

3.4.3 Assessing the predictive performances

Finally, the predictive performance of the model is assessed. In addition to the non-separable model according to which data are simulated, we also consider three more specific models, which are special cases of the full model. We compare space-time separable (S) and non-separable (NS) models, and models with equal (E) or different (D) smoothness and scale parameters. We will thus consider the following 4 models :

1. S-E : $b = 0$; all components have equal smoothness and scale parameters. The covariance matrix follows (3.2).

	σ_1		σ_2		σ_3		β_{12}		β_{13}		β_{23}	
	WPL	FL	WPL	FL	WPL	FL	WPL	FL	WPL	FL	WPL	FL
$bs^2 (\times 10^{-5})$	0.18	5.13	1.13	2.54	3.48	17.1	1.02	5.77	6.30	4.50	7.96	5.35
$sd^2 (\times 10^{-3})$	1.73	0.84	1.40	0.59	1.24	0.63	8.47	0.26	7.86	0.29	10.8	0.36
rmse	0.042	0.030	0.038	0.025	0.036	0.028	0.092	0.018	0.089	0.018	0.104	0.020
rre	0.72		0.66		0.79		0.19		0.21		0.19	
	v_1		v_2		v_3		α		a		b	
	WPL	FL	WPL	FL	WPL	FL	WPL	FL	WPL	FL	WPL	FL
$bs^2 (\times 10^{-3})$	2.63	3.48	1.75	6.30	0.60	2.95	0.132	2.82	0.001	0.018	1.000	27.0
$sd^2 (\times 10^{-3})$	9.71	4.30	13.1	8.95	3.81	1.82	11.4	4.70	5.08	0.699	52.7	19.4
rmse	0.111	0.088	0.122	0.123	0.066	0.069	0.107	0.087	0.071	0.027	0.232	0.215
rre	0.79		1.01		1.04		0.81		0.38		0.93	
	$1/r_1$		$1/r_2$		$1/r_3$							
	WPL	FL	WPL	FL	WPL	FL						
bs^2	218.5	911.6	61.8	473.9	516.6	4539.3						
sd^2	2300.0	856.1	1127.1	682.4	9629	6788.9						
rmse	50.2	42.0	34.5	34.0	100.7	106.4						
rre	0.84		0.99		1.06							

TABLE 3.4 – Bias (bs), variance (sd^2), root mean square error (rmse) and root relative efficiency (rre) for each parameter. Same simulations as in Table 3.1.

2. NS-E : $0 < b \leq 1$; all components have equal smoothness and scale parameters. This model is space-time non-separable, but it has the same space-time covariance for all variables. The covariance matrix-valued function follows (3.1).
3. S-D : $b = 0$; each component has its own smoothness and scale parameters, but it is space-time separable.
4. NS-D : $0 < b \leq 1$; each component has its own smoothness and scale parameters, and it is not space-time separable. It is the non-separable model defined through Eq. (3.12).

Data are simulated according to model NS-D, with parameters as in Table 3.3, at the 13 stations displayed in Figure 3.4. Two stations, indicated with a star, have been selected for validation. Data at these two stations are not used to estimate the parameters. For each day t , the conditional distribution is computed at these two locations, given the sets $\{\text{all data at time } t-1, \dots, t-q\} \cup \{\text{data at other locations at the same time } t\}$, where $q \in \{2, 3, \dots\}$ is the number of days used for the prediction. Under Gaussianity, this amounts to compute the 6-variate conditional expectation and the conditional covariance matrix.

The four models are compared by means of four different scores : Root Mean Square Error (RMSE), Mean Absolute Error (MAE), the Continuous Ranked Probability Score (CRPS) and Logarithmic Score (LogS) [Gneiting and Raftery, 2007]. Let us denote $\tilde{z}_1^t, \dots, \tilde{z}_6^t$ the conditional expectation of the 6 predicted variables (2 sites $\times$ 3 variables) for a given day t , $\tilde{\sigma}_1, \dots, \tilde{\sigma}_6$ the corresponding conditional standard deviations, $z_1^t, \dots, z_6^t$ being the observed values. Note that the conditional standard deviations are independent of the day t , since they only depend on covariance values.

The first two scores, MAE and RMSE, compare the conditional expectation to the true value. The MAE is defined as

$$\text{MAE} = \frac{1}{6|T|} \sum_{t \in T} \sum_{j=1}^6 |\tilde{z}_j^t - z_j^t|,$$

where the sum is taken over the set of all testing days, T and $|T|$ denotes the cardinality of the set T . The mean square error (MSE) is

$$\text{MSE} = \frac{1}{6|T|} \sum_{t \in T} \sum_{j=1}^6 (\tilde{z}_j^t - z_j^t)^2.$$

The RMSE is the square root of the MSE and has the advantage of being recorded in the same unit as the data.

The other two scores, CRPS and LogS, assess not only the prediction but its variance as well. They are easily computed in the case of a normal predictive distribution. The CRPS measures the discrepancy between the predictive cumulative distribution function (CDF) and the true value. Specifically, if F is the predictive CDF and z the true value, crps is defined as

$$\text{crps}(F, z) = \int_{-\infty}^{\infty} [F(u) - H(u - z)]^2 du,$$

where $H(u - z)$ denotes the Heaviside function, which takes the value 0 when $u < z$ and 1 otherwise. In case of normal CDF Φ_j^t with expectation $\tilde{z}_j^t$ and variance $\tilde{\sigma}_j^2$, one can show that

$$\text{crps}(\Phi_j^t, z) = \tilde{\sigma}_j^t \{ \tilde{z} [2\Phi(\tilde{z}) - 1] + 2\phi(\tilde{z}) - \pi^{-1/2} \},$$

where $\tilde{z} = (z - \tilde{z}_j^t) / \tilde{\sigma}_j$ is the normalized prediction error and where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and the CDF, respectively, of the normal distribution with mean 0 and variance 1. The CRPS is

$$\text{CRPS} = \frac{1}{6|T|} \sum_{t \in T} \sum_{j=1}^6 \text{crps}(\Phi_j^t, y_j^t).$$

It is easy to show that the CRPS tends to the MAE when $\tilde{\sigma}_j \rightarrow 0$, for all $j = 1, \dots, 6$. For this reason, the CRPS can be interpreted as a generalized version of the MAE [Gneiting and Raftery, 2007]. The marginal logarithmic score is the negative of the logarithm of the marginal predictive density at the true value

$$\text{LogS}_1 = \frac{1}{6|T|} \sum_{t \in T} \sum_{j=1}^6 \left\{ \ln \tilde{\sigma}_j + \frac{(\tilde{z}_j^t - z_j^t)^2}{2\tilde{\sigma}_j^2} \right\}.$$

The multivariate logarithmic score considers the multivariate predictive density computed at the true vector $\mathbf{z}^t = (z_1, \dots, z_6)^\top$

$$\text{LogS}_6 = \frac{1}{6|T|} \sum_{t \in T} \left\{ \ln \det \tilde{\Sigma}^t + \frac{(\tilde{\mathbf{z}}^t - \mathbf{z}^t)^\top \tilde{\Sigma}^{-1} (\tilde{\mathbf{z}}^t - \mathbf{z}^t)}{2} \right\},$$

where $\tilde{\Sigma}$ is the conditional covariance matrix and $\mathbf{a}^\top$ is the transpose of $\mathbf{a}$.

The scores defined above are computed for the set of 100 simulations, for each of the four models, and for both estimation approaches, WPL and FL. The average scores are reported in Table 3.5, and the boxplots for WPL are shown in Figure 3.3. We only report results obtained using the two previous days as conditioning data (i.e. $q = 2$), but similar results were obtained for $q = 1$ and $q = 3$. Recall that lower scores indicate a better adequacy between the model and the data.

Some comments are in order : in general, differences in score values are low, but they are consistently observed on all simulations. Therefore, they can be considered as significant. Obviously, the highest scores correspond to the fully separable model, SE, whereas the lowest scores are obtained for the non-separable model, NS-D, for which the difference between WPL and FL is negligible. Differences between models are more pronounced when considering scores that also involve the conditional variance (CRPS and LogS). Accounting for non-separability makes a more important difference than having different smoothness and scale parameters for each variable. It is interesting to note that the difference between the models is more pronounced with WPL than with FL.

In conclusion, this simulation study shows that WPL is a valid procedure for estimating the parameters of the Gneiting-Matérn multivariate space-time model when analyzing large data sets for which a Full Likelihood maximization is not feasible. The fact that

	Weighted Pairwise Likelihood					Full Likelihood				
	RMSE	MAE	CRPS	LogS ₁	LogS ₆	RMSE	MAE	CRPS	LogS ₁	LogS ₆
S-E	0.494	0.391	0.278	0.725	4.070	0.489	0.386	0.274	0.690	3.792
NS-E	0.488	0.386	0.274	0.701	3.903	0.488	0.386	0.273	0.688	3.782
S-D	0.490	0.389	0.277	0.729	4.091	0.484	0.383	0.271	0.680	3.733
NS-D	0.485	0.383	0.271	0.683	3.789	0.484	0.383	0.271	0.678	3.723

TABLE 3.5 – Mean of the prediction scores RMSE, MAE, CRPS, LogS₁ and LogS₆, according to models S-E, NS-E, S-D and NS-D, and for WPL and FL approaches. Same simulations as in Table 3.1.

the difference in scores between the models is more pronounced with WPL than with FL is a strong indication that it is particularly interesting to apply WPL to the non-separable NS-D model. When selecting the correct model, we note that in terms of prediction performances (RMSE, MAE, CRPS), the differences are negligible when using WPL estimates instead of FL estimates despite the lower estimating efficiency shown in Table 3.4.

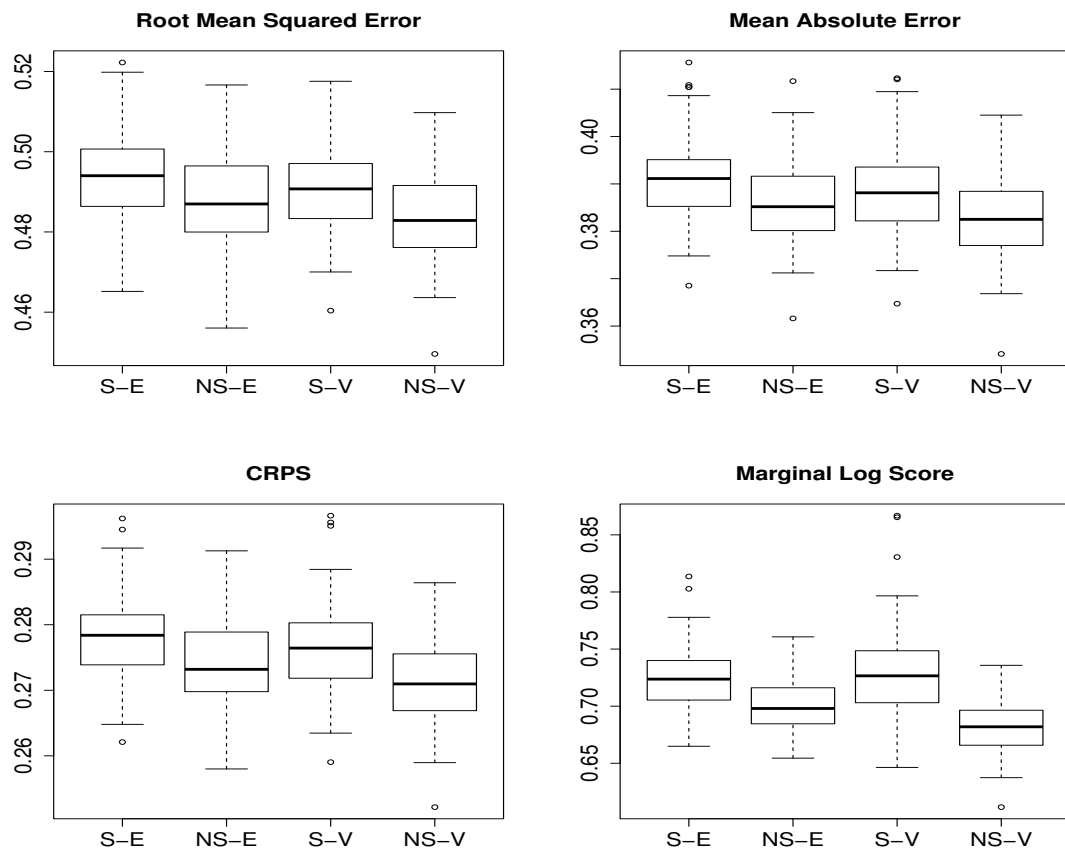


FIGURE 3.3 – Prediction scores RMSE, MAE, CRPS and LogS_1 for the models S-E, NS-E, S-D and NS-D. Same simulations as in Table 3.1. Estimates obtained with WPL.

3.5 Western France weather dataset

This section illustrates the use of the Gneiting-Matérn multivariate space-time model for the analysis of a weather dataset which consists of three daily variables (solar radiation, R , temperature, T , and humidity, H) recorded at 13 stations in Western France from 2003 to 2012 (see Figure 3.4). These data are part of the French National Institute for Agricultural Research (INRA) archive of weather data. The considered domain experiences an oceanic climate, characterized by moist and cool (but not cold) winters. Due to the prevailing westerly winds, some amount of space-time interaction is thus expected, from West to East. The validation stations (indicated with stars in Figure 3.4) have been selected in the Eastern part of the domain, neither too close, nor too far from the other stations. The first validation station, Le Rheu, is located near Rennes, in Brittany. The second one, Bourran, is located near Agen, in Aquitaine. The other 11 stations are used to estimate the parameters. In order to restrict ourselves to data being stationary in time, we selected data recorded in January, from 2003 to 2012.

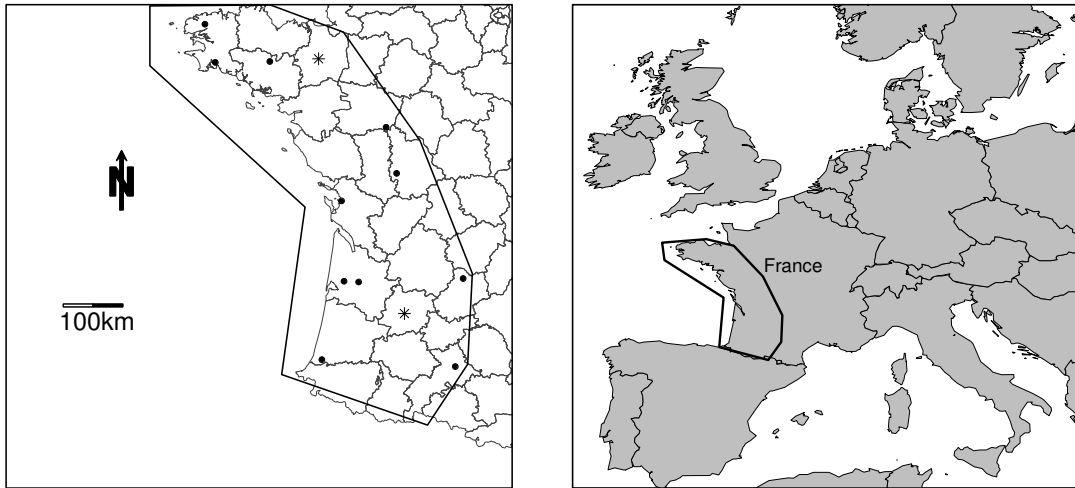


FIGURE 3.4 – Location of the 13 weather stations over western France. Stars : validation stations.

To model the January data $\mathbf{Y}(\mathbf{s}, t) = \{Y_R(\mathbf{s}, t), Y_T(\mathbf{s}, t), Y_H(\mathbf{s}, t)\}^\top$, we consider a model that is stationary in time and non-stationary in space, namely

$$Y_i(\mathbf{s}, t) = \mu_i(\mathbf{s}) + \sigma_i(\mathbf{s})Z_i(\mathbf{s}, t) \quad i \in \{R, T, H\} \quad (\mathbf{s}, t) \in \mathbb{R}^2 \times \mathbb{R}. \quad (3.13)$$

The data are first standardized at each location by their averages and standard deviations. The space-time modeling will be carried out on the standardized variables,

$$\mathbf{Z}(\mathbf{s}, t) = \{Z_R(\mathbf{s}, t), Z_T(\mathbf{s}, t), Z_H(\mathbf{s}, t)\}^\top,$$

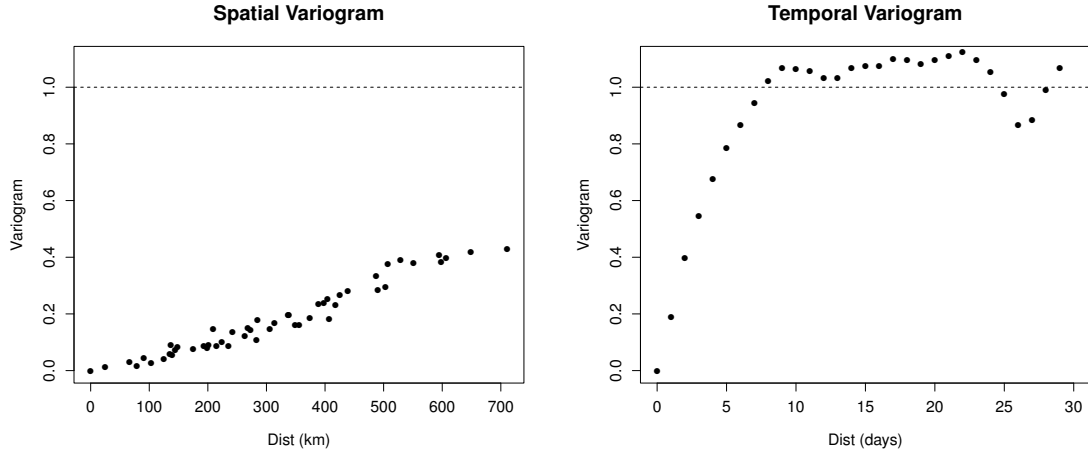


FIGURE 3.5 – Marginal spatial and temporal variograms for daily January temperature standardized variables.

which are assumed to be space-time stationary and multivariate Gaussian. The simultaneous observation of the spatial and temporal variograms of the daily temperature standardized variables, depicted in Figure 3.5, reveals a zonal anisotropy between space and time. While the temporal variogram reaches a sill equal to the overall variance, equal to 1 due to the standardization, the spatial variogram does not exceed 0.4, even at very large distances. This zonal anisotropy is modeled by considering that the process $\mathbf{Z}$ arises as the sum of a zero mean multivariate temporal Gaussian process $\mathbf{X}$ accounting for temporal effects being constant in space, and an independent zero mean multivariate space-time Gaussian process $\mathbf{W}$

$$Z_i(\mathbf{s}, t) = X_i(t) + W_i(\mathbf{s}, t) \quad i \in \{\mathbf{R}, \mathbf{T}, \mathbf{H}\}, \quad (\mathbf{s}, t) \in \mathbb{R}^2 \times \mathbb{R}. \quad (3.14)$$

Physically speaking, the model (3.14) assumes that part of the daily changes in the weather pattern, which accounts for 60% of the overall variability, is regional and affects the whole domain under study. The stationary space-time variations account for 40% of the variability.

This model provides the flexibility to accurately model the zonal anisotropy observed in Figure 3.5, since

$$\begin{aligned} \gamma_{ii}^{(\mathbf{Z})}(\mathbf{h}, 0) &= \gamma_{ii}^{(\mathbf{X})}(0) + \gamma_{ii}^{(\mathbf{W})}(\mathbf{h}, 0) \xrightarrow{\mathbf{h} \rightarrow \infty} \left[\sigma_i^{(\mathbf{W})} \right]^2 \\ \gamma_{ii}^{(\mathbf{Z})}(\mathbf{0}, u) &= \gamma_{ii}^{(\mathbf{X})}(u) + \gamma_{ii}^{(\mathbf{W})}(\mathbf{0}, u) \xrightarrow{u \rightarrow \infty} \left[\sigma_i^{(\mathbf{X})} \right]^2 + \left[\sigma_i^{(\mathbf{W})} \right]^2 = 1. \end{aligned}$$

We further suppose that the process $\mathbf{W}$ is Gaussian, with a Gneiting-Matérn covariance function, as described in Eq. (3.6) and that the temporal process $\mathbf{X}$ has covariance function $C_{ij}^{(\mathbf{X})}(u) = \sigma_i^{(\mathbf{X})} \sigma_j^{(\mathbf{X})} \beta_{ij}^{(\mathbf{X})} / (\alpha^{(\mathbf{X})} |u|^{2a^{(\mathbf{X})}} + 1)$, with $i, j \in \{\mathbf{R}, \mathbf{T}, \mathbf{H}\}$ and $u \in \mathbb{R}$. Note that this temporal covariance function is the temporal marginal of a Gneiting-Matérn covariance (i.e. when $\mathbf{h} = \mathbf{0}$).

	$\sigma_R^{(X)}$	$\sigma_T^{(X)}$	$\sigma_H^{(X)}$	$\beta_{RT}^{(X)}$	$\beta_{RH}^{(X)}$	$\beta_{TH}^{(X)}$	$\sigma_R^{(W)}$	$\sigma_T^{(W)}$	$\sigma_H^{(W)}$	$\beta_{RT}^{(W)}$	$\beta_{RH}^{(W)}$	$\beta_{TH}^{(W)}$
S-E	0.38	0.90	0.49	-0.72	-0.52	0.66	0.93	0.43	0.87	-0.46	-0.45	-0.02
NS-E	0.33	0.90	0.45	-0.79	-0.52	0.71	0.94	0.44	0.89	-0.47	-0.45	-0.02
S-D	0.20	0.85	0.43	-0.98	-0.56	0.72	0.98	0.53	0.90	-0.53	-0.44	0.08
NS-D	0.19	0.87	0.44	-0.97	-0.55	0.73	0.98	0.50	0.90	-0.59	-0.47	0.06
	v_R	v_T	v_H	$1/r_R$	$1/r_T$	$1/r_H$	$\alpha^{(X)}$	$a^{(X)}$	$\alpha^{(W)}$	$a^{(W)}$	b	-wcl
S-E	0.66	0.66	0.66	184	184	184	0.12	1.00	2.35	0.89	0	631995
NS-E	0.67	0.67	0.67	188	188	188	0.13	0.93	2.08	1.00	0.6	631991
S-D	0.75	0.84	0.41	211	249	248	0.09	1.00	0.91	0.69	0	631922
NS-D	0.64	0.84	0.30	239	224	322	0.12	1.00	1.06	0.75	0.6	631845

TABLE 3.6 – Estimates of the parameters using WPL with $\mathbf{d} = (500\text{km}, 2\text{days})$. Unit is kms for distances.

Regarding the process $\mathbf{W}$, we consider the same four space-time models as in the previous section, for which there are at most 15 parameters to be estimated (for model NS-D). There are 8 parameters for the process $\mathbf{X}$. Since the data have been standardized, we impose $\left[\sigma_i^{(X)}\right]^2 + \left[\sigma_i^{(W)}\right]^2 = 1$, $i \in \{R, T, H\}$, thus leading to a total of at most 20 parameters. Estimates are reported in Table 3.6. We also find that the model NS-E performs worst than S-E. Figures 3.6 and 3.7 represent the empirical spatial direct and cross-covariance functions and the model NS-D with estimated parameters using WPL. We observe that, in general, the model fits quite well the experimental covariance. In particular, the model is able to fit the different regularities, as well as the space-time interactions. It is slightly better fitted for the spatial dimensions than along time, probably because of much less flexibility in the model along time. Conditional expectation and envelopes of 100 conditional simulations are shown in Figure 3.8 at the two validation locations. The true values are always contained within the envelopes.

The analysis of the colocated covariance coefficients between the three variables reveals an interesting pattern. We compute the estimated covariance coefficient $\sigma_{ij}^{(X)} = \sigma_i^{(X)} \sigma_j^{(X)} \rho_{ij}^{(X)}$, with $i, j \in \{R, T, H\}$, where $\rho_{ij}^{(X)}$ is the correlation coefficient as defined in Theorem 1. Similar coefficients are computed for $\mathbf{W}$ and $\mathbf{Z}$ and compared to the empirical correlation coefficients in Table 3.8. The correlation coefficients estimated for $\mathbf{Z}$ are very close to empirical ones. Note that Radiation is always negatively correlated to Temperature and Humidity, which is typical for winter weather conditions. More interestingly, we can observe that the (negative) correlation between Radiation and Humidity is mostly due to the space-time process $\mathbf{W}$, whereas the (positive) correlation between Temperature and Humidity is mostly due to temporal process $\mathbf{X}$.

	RMSE	MAE	CRPS	LogS ₁
S-E	0.433	0.306	0.222	0.440
NS-E	0.446	0.313	0.228	0.467
S-D	0.417	0.295	0.215	0.405
NS-D	0.417	0.295	0.216	0.397

TABLE 3.7 – RMSE, MAE, CRPS and LogS₁ of predicted values at the validation stations, using WPL estimates with $\mathbf{d} = (500\text{km}, 2\text{days})$.

	σ_{RT}	σ_{RH}	σ_{TH}
Process X	−0.16	−0.05	0.28
Process W	−0.29	−0.39	0.03
Process Z	−0.45	−0.44	0.31
Empirical	−0.38	−0.42	0.25

TABLE 3.8 – Empirical and estimated covariance coefficients $\sigma_{ij} = \sigma_i \sigma_j \rho_{ij}$, with $i, j \in \{R, T, H\}$ for the model NS-D.

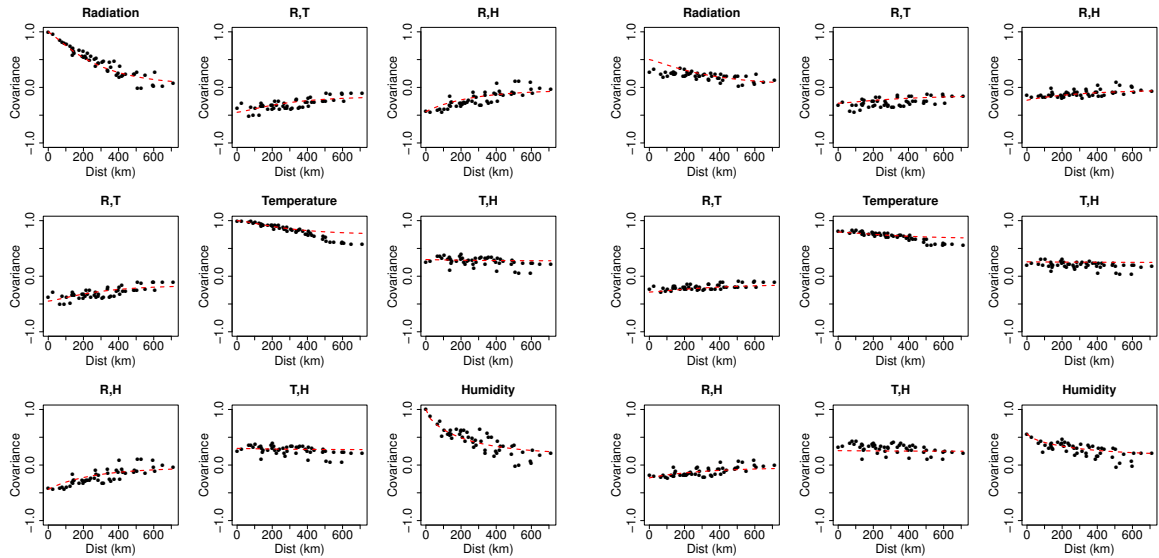


FIGURE 3.6 – Spatial direct and cross-covariance functions for R, T and H. Points : empirical values. Solid line : model NS-D with parameters estimated using WPL (see Table 3.6). Left panel : $u = 0$. Right panel : $u = 1$.

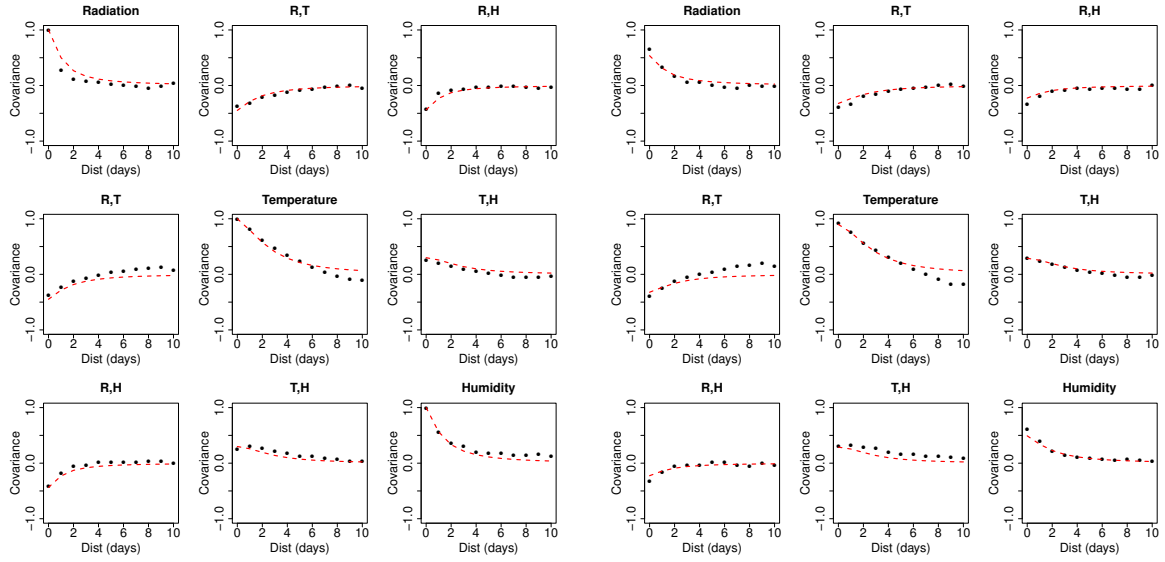


FIGURE 3.7 – Temporal direct and cross-covariance functions for R, T and H, as in Figure 3.6. Left panel : $\mathbf{h} = 0$. Right panel : $\|\mathbf{h}\| = 199$ km.

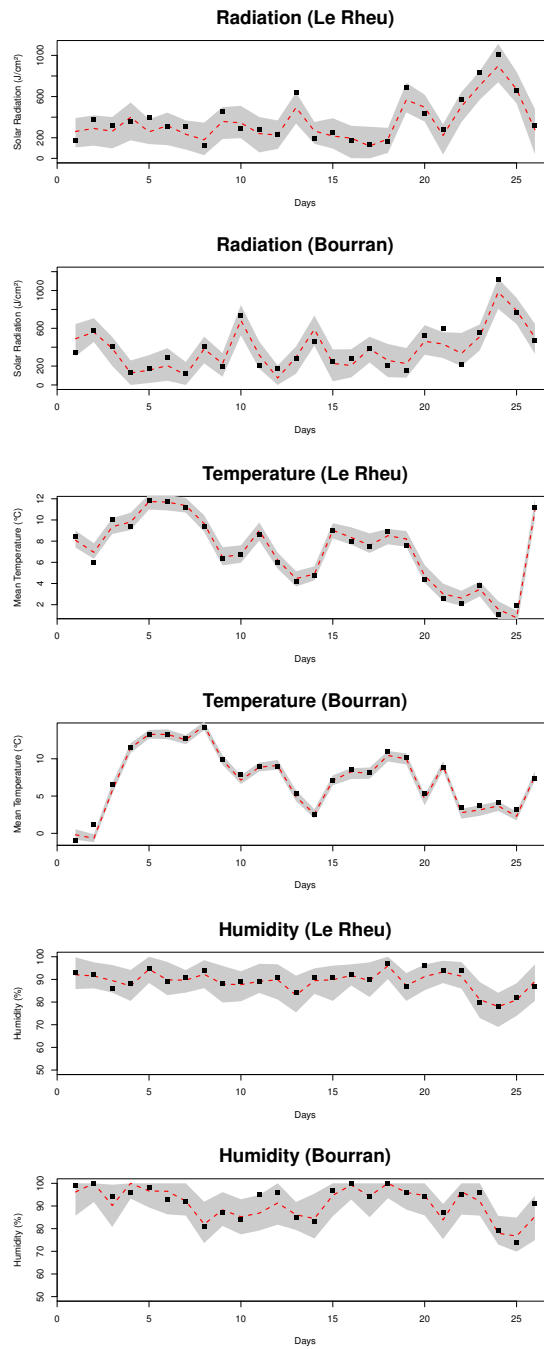


FIGURE 3.8 – Prediction of Radiation, Temperature and Humidity at the two validation stations, Le Rheu (Brittany) and Bourran (Aquitaine). Black points : true values. Dotted lines : conditional expectation. Shaded area : 90% envelope interval, computed on 100 simulations.

3.6 Discussion

Flexible multivariate space-time models can be built using a mixture approach. We proposed two classes based on the univariate Gneiting class of space-time covariances : the Gneiting-Matérn and the Gneiting-Cauchy multivariate space-time models. A simulation study showed that the relatively large number of parameters could be accurately estimated by maximizing a weighted pairwise likelihood function. Whenever data are simulated from the most general model, the analysis of the estimation bias and variance, as well as that of validation scores revealed that ignoring space-time interaction or the multivariate flexibility consistently leads to non optimal scores. The relative efficiency of estimates obtained when maximizing weighted pairwise likelihood with respect to those obtained when maximizing a full likelihood is, in general, above 70%. It thereby confirms that weighted pairwise likelihood is a reliable alternative to full likelihood when analyzing large datasets.

Our approach could be generalized in several ways. First, with similar arguments to those used in Theorems 1 and 2, we could propose a wider range of flexible multivariate space-time models, including compactly supported models, and models on manifolds such as spheres. One limit of our models is that there is a unique temporal covariance for all variables. By exchanging the role of time and space, one can easily build models that are flexible in time and unique in space. Building valid multivariate space-time models that are flexible both in time and space is still an open challenge. For the time being, the analysis of models requiring different temporal covariances for different variables can be carried out using a sum of models, as in Section 3.5.

WPL has been shown to provide accurate estimates for univariate spatial and space-time Gaussian random fields [Bevilacqua and Gaetan, 2015; Bevilacqua et al., 2012]. We showed that it also provides accurate estimates for multivariate space-time Gaussian random fields. We found, however, that the space-time separability parameter, b , was particularly difficult to estimate. The effect of the separability, which could be formally defined as the ratio between the non-separable space-covariance and the separable one, decreases as $\|\mathbf{h}\| \rightarrow 0$ or $u \rightarrow 0$. Therefore, the relative efficiency of the estimator of b decreases when the window $\mathbf{d} = (d_S, d_T)$ defining the pairs in WPL is too small. The optimal size of this window is thus a balance between two opposite requirements : accurately estimating b , while offering speed gain. Our opinion is that one could probably increase the efficiency of the estimation of the parameters without sacrificing too much computation speed by proposing new forms of composite likelihoods.

Appendix

Our proofs are based on the scale mixture representation (3.5), for which, according to Lemma 1, we need the following ingredients : a relevant univariate covariance function $C_\xi(\mathbf{h}, u)$ in $\mathbb{R}^k \times \mathbb{R}$ and a non-negative definite multivariate mixture $\mathbf{m}(\xi) = [m_{ij}(\xi)]_{i,j=1}^p$, $\xi > 0$. The measure F in Lemma 1 will be set to be the Lebesgue measure on $(0, \infty)$.

The first part of the proof is common to both Theorems. We first verify that the function

$$C_\xi(\mathbf{h}, u) = \frac{1}{\psi(u^2)^{k/2}} \exp \left\{ -\xi \left(\frac{\|\mathbf{h}\|}{\psi(u^2)^{1/2}} \right)^\lambda \right\}, \quad (3.15)$$

defined on $\mathbb{R}^k \times \mathbb{R}$ with $0 < \lambda \leq 2$, is a valid covariance function for any $\xi > 0$ and for $\psi(t), t \geq 0$, being a positive function with a completely monotone derivative. Indeed, the mapping $t \mapsto \exp(-ct^\gamma)$, $t \geq 0$, $c > 0$ and $0 < \gamma \leq 1$ is completely monotone. Then, setting $\gamma = 1$, $c = \left(\frac{\|\mathbf{h}\|}{\psi(u^2)^{1/2}} \right)^\lambda$ and $t = \xi$, it is clear that $C_\xi(\mathbf{h}, u)$ belongs to the Gneiting class of covariance functions [Gneiting, 2002] for any $\xi > 0$. Also, the mapping $[0, \infty) \rightarrow \mathbb{R} : \xi \mapsto C_\xi(\mathbf{h}, u)$ satisfies the hypothesis of Lemma 1, for all fixed $(\mathbf{h}, u) \in \mathbb{R}^k \times \mathbb{R}$.

The rest of the proofs is splitted into two parts corresponding to each model. It consists in finding multivariate mixtures $\mathbf{m}(\xi) = [m_{ij}(\xi)]_{i,j=1}^p \in \mathbb{R}^p \times \mathbb{R}^p$ tailored to either the Matérn or the Cauchy class, that are symmetric and non-negative definite for all $\xi \geq 0$.

Proof of Theorem 1

Let $\nu > 0$ and $r > 0$. From Eq. 3.471.9 in Gradshteyn and Ryzhik [2007], we have

$$\mathcal{M}(\mathbf{h}; r, \nu) = \int_0^\infty \exp\{-\|\mathbf{h}\|^2 \xi\} m^{\mathcal{M}}(\xi; r, \nu) d\xi,$$

with

$$m^{\mathcal{M}}(\xi; r, \nu) = \left(\frac{r^2}{4} \right)^\nu \frac{\xi^{-1-\nu}}{\Gamma(\nu)} \exp \left\{ -\frac{r^2}{4\xi} \right\}, \quad \xi > 0.$$

Following the univariate mixture representation above, we set

$$\mathbf{m}^{\mathcal{M}}(\xi) = [m_{ij}^{\mathcal{M}}(\xi)]_{i,j=1}^p = \left[\rho_{ij} \left(\frac{r_{ij}^2}{4} \right)^{\nu_{ij}} \frac{\xi^{-1-\nu_{ij}}}{\Gamma(\nu_{ij})} \exp \left\{ -\frac{r_{ij}^2}{4\xi} \right\} \right]_{i,j=1}^p, \quad \xi > 0,$$

with

$$\begin{aligned} r_{ij} &= \{(r_i^2 + r_j^2)/2\}^{1/2}, \quad \forall i, j = 1, \dots, p \text{ and } r_i > 0, \quad \forall i = 1, \dots, p, \\ \nu_{ij} &= (\nu_i + \nu_j)/2, \quad \forall i, j = 1, \dots, p, \\ \rho_{ij} &= \beta_{ij} \frac{\Gamma(\nu_{ij})}{\Gamma(\nu_i)^{1/2} \Gamma(\nu_j)^{1/2}} \frac{r_i^{\nu_i} r_j^{\nu_j}}{r_{ij}^{2\nu_{ij}}}, \quad \forall i, j = 1, \dots, p, \end{aligned}$$

where $\boldsymbol{\beta} = [\beta_{ij}]_{i,j=1}^p$ is a correlation matrix. We have, for each $i, j = 1, \dots, p$,

$$\begin{aligned} m_{ij}^{\mathcal{M}}(\xi) &= \rho_{ij} \left(\frac{r_{ij}^2}{4} \right)^{v_{ij}} \frac{\xi^{-1-v_{ij}}}{\Gamma(v_{ij})} \exp \left\{ -\frac{r_{ij}^2}{4\xi} \right\} \\ &= \beta_{ij} \frac{\Gamma(v_{ij})}{\Gamma(v_i)^{1/2} \Gamma(v_j)^{1/2}} \frac{r_i^{v_i} r_j^{v_j}}{r_{ij}^{2v_{ij}}} \left(\frac{r_{ij}^2}{4} \right)^{v_{ij}} \frac{\xi^{-1-v_{ij}}}{\Gamma(v_{ij})} \exp \left\{ -\frac{r_{ij}^2}{4\xi} \right\} \\ &= \beta_{ij} \frac{r_i^{v_i}}{2^{v_i} \Gamma(v_i)^{1/2}} \xi^{-(1+v_i)/2} \exp \left\{ -\frac{r_i^2}{8\xi} \right\} \cdot \frac{r_j^{v_j}}{2^{v_j} \Gamma(v_j)^{1/2}} \xi^{-(1+v_j)/2} \exp \left\{ -\frac{r_j^2}{8\xi} \right\} \\ &= \beta_{ij} m_i^{\mathcal{M}}(\xi)^{1/2} m_j^{\mathcal{M}}(\xi)^{1/2}. \end{aligned}$$

Therefore $\mathbf{m}^{\mathcal{M}}(\xi)$ is the Hadamard product of the correlation matrix $\boldsymbol{\beta}$ and the outer product $\bar{\mathbf{m}}^{\mathcal{M}}(\xi)$ by itself, with $\bar{\mathbf{m}}^{\mathcal{M}}(\xi) = (m_1^{\mathcal{M}}(\xi)^{1/2}, \dots, m_p^{\mathcal{M}}(\xi)^{1/2})^\top$. By Schur's Theorem [Horn and Johnson, 2012, p. 455], the matrix $\mathbf{m}^{\mathcal{M}}(\xi)$ is thus non-negative definite since it is the Hadamard product of two non-negative definite matrices.

Setting $\lambda = 2$ in Eq. (3.15) with $\mathbf{m}^{\mathcal{M}}(\xi) = [m_{ij}^{\mathcal{M}}(\xi)]_{i,j=1}^p$ as defined above leads thus to the valid p -variate matrix-valued covariance function $[C_{ij}^{\mathcal{M}}(\mathbf{h}, u)]_{i,j=1}^p$ on $\mathbb{R}^k \times \mathbb{R}$ by application of Lemma 1 :

$$\begin{aligned} C_{ij}^{\mathcal{M}}(\mathbf{h}, u) &= \sigma_i \sigma_j \int_0^\infty \frac{1}{\psi(\|\mathbf{u}\|^2)^{k/2}} \exp \left\{ -\xi \left(\frac{\|\mathbf{h}\|}{\psi(u^2)^{1/2}} \right)^2 \right\} m_{ij}^{\mathcal{M}}(\xi) d\xi \\ &= \frac{\sigma_i \sigma_j}{\psi(u^2)^{k/2}} \rho_{ij} \mathcal{M} \left(\frac{\mathbf{h}}{\psi(\|\mathbf{u}\|^2)^{1/2}}; r_{ij}, v_{ij} \right) \end{aligned}$$

□

Proof of Theorem 2

It is well known that the Cauchy covariance function is the Laplace transform of a Gamma distribution. Therefore, we set

$$\mathbf{m}^{\mathcal{C}}(\xi) = [m_{ij}^{\mathcal{C}}(\xi)]_{i,j=1}^p = \left[\rho_{ij} \frac{1}{r_{ij}^{v_{ij}}} \frac{\xi^{v_{ij}-1}}{\Gamma(v_{ij})} \exp \left\{ -\frac{1}{r_{ij}} \xi \right\} \right]_{i,j=1}^p, \quad \xi > 0,$$

with

$$\begin{aligned} r_{ij} &= \{(r_i^{-1} + r_j^{-1})/2\}^{-1}, \quad \forall i, j = 1, \dots, p \text{ and } r_i > 0, \quad \forall i = 1, \dots, p, \\ v_{ij} &= (v_i + v_j)/2, \quad \forall i, j = 1, \dots, p, \\ \rho_{ij} &= \beta_{ij} \frac{\Gamma(v_{ij})}{\Gamma(v_i)^{1/2} \Gamma(v_j)^{1/2}} \frac{r_{ij}^{v_{ij}}}{(r_i^{v_i} r_j^{v_j})^{1/2}}, \quad \forall i, j = 1, \dots, p, \end{aligned}$$

where $\boldsymbol{\beta} = [\beta_{ij}]_{i,j=1}^p$ is a correlation matrix. This multivariate mixture is non-negative de-

finite. Indeed, we have

$$\begin{aligned}
 m_{ij}^{\mathcal{C}}(\xi) &= \rho_{ij} \frac{1}{r_{ij}^{v_{ij}}} \frac{\xi^{v_{ij}-1}}{\Gamma(v_{ij})} \exp\left\{-\frac{1}{r_{ij}}\xi\right\} \\
 &= \beta_{ij} \frac{\Gamma(v_{ij})}{\Gamma(v_i)^{1/2}\Gamma(v_j)^{1/2}} \frac{r_{ij}^{v_{ij}}}{(r_i^{v_i} r_j^{v_j})^{1/2}} \frac{1}{r_{ij}^{v_{ij}}} \frac{\xi^{v_{ij}-1}}{\Gamma(v_{ij})} \exp\left\{-\frac{1}{r_{ij}}\xi\right\} \\
 &= \beta_{ij} \frac{\xi^{\frac{v_i-1}{2}}}{(\Gamma(v_i) r_i^{v_i})^{1/2}} \exp\left\{-\frac{1}{2r_i}\xi\right\} \cdot \frac{\xi^{\frac{v_j-1}{2}}}{(\Gamma(v_j) r_j^{v_j})^{1/2}} \exp\left\{-\frac{1}{2r_j}\xi\right\} \\
 &= \beta_{ij} m_i^{\mathcal{C}}(\xi)^{1/2} m_j^{\mathcal{C}}(\xi)^{1/2}.
 \end{aligned}$$

The same arguments as in Theorem 1 are used to conclude that $\mathbf{m}^{\mathcal{C}}(\xi)$ is non-negative definite. The last step of the proof consists in applying Lemma 1 with $\mathbf{m}^{\mathcal{C}}(\xi) = \left[m_{ij}^{\mathcal{C}}(\xi)\right]_{i,j=1}^p$ as defined above and by setting $\lambda = 2$ in Eq. (3.15). The result is the valid p -variate matrix-valued covariance function on $\mathbb{R}^k \times \mathbb{R}$

$$\begin{aligned}
 C_{ij}^{\mathcal{C}}(\mathbf{h}, u) &= \sigma_i \sigma_j \int_0^\infty \frac{1}{\psi(u^2)^{k/2}} \exp\left\{-\xi \left(\frac{\|\mathbf{h}\|}{\psi(u^2)^{1/2}}\right)^\lambda\right\} m_{ij}^{\mathcal{C}}(\xi) d\xi \\
 &= \frac{\sigma_i \sigma_j}{\psi(u^2)^{k/2}} \rho_{ij} \int_0^\infty \exp\left\{-\xi \left(\frac{\|\mathbf{h}\|}{\psi(u^2)^{1/2}}\right)^\lambda\right\} \frac{1}{r_{ij}^{v_{ij}}} \frac{\xi^{v_{ij}-1}}{\Gamma(v_{ij})} \exp\left\{-\frac{1}{r_{ij}}\xi\right\} d\xi \\
 &= \frac{\sigma_i \sigma_j}{\psi(u^2)^{k/2}} \rho_{ij} \left\{1 + r_{ij} \left(\frac{\|\mathbf{h}\|}{\psi(u^2)^{1/2}}\right)^\lambda\right\}^{-v_{ij}} \\
 &\quad \times \int_0^\infty \left\{\frac{1}{r_{ij}} + \left(\frac{\|\mathbf{h}\|}{\psi(u^2)^{1/2}}\right)^\lambda\right\}^{v_{ij}} \frac{\xi^{v_{ij}-1}}{\Gamma(v_{ij})} \exp\left\{-\xi \left[\frac{1}{r_{ij}} + \left(\frac{\|\mathbf{h}\|}{\psi(u^2)^{1/2}}\right)^\lambda\right]\right\} d\xi \\
 &= \frac{\sigma_i \sigma_j}{\psi(u^2)^{k/2}} \rho_{ij} \mathcal{C}\left(\frac{\mathbf{h}}{\psi(u^2)^{1/2}}; r_{ij}, v_{ii}, \lambda\right) \square
 \end{aligned}$$

Acknowledgments

Two anonymous referees are gratefully acknowledged. The authors wish to thank Carlo Gaetan from the Ca' Foscari University in Venezia (Italy) for fruitful discussions and Liliane Bel from AgroParisTech in Paris (France) for her very careful reading of earlier versions of this manuscript. This work was supported by the metaprogramme Adaptation of Agriculture and Forests to Climate Change (AAFCC) of the French National Institute for Agricultural Research (INRA). Emilio Porcu has been supported by Proyecto Fondecyt n. 1130647 from Chilean Ministry of Education.

3.7 Références

Apanasovich, T. V. and Genton, M. G. (2010). Cross-covariance functions for multivariate random fields based on latent dimensions. *Biometrika*, 97(1) :15–30. [43](#)

- Apanasovich, T. V., Genton, M. G., and Sun, Y. (2012). A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components. *Journal of the American Statistical Association*, 107(497) :180–193. [42](#), [43](#), [44](#), [46](#)
- Bevilacqua, M. and Gaetan, C. (2015). Comparing composite likelihood methods based on pairs for spatial Gaussian random fields. *Statistics and Computing*, 25(5) :877–892. [47](#), [64](#)
- Bevilacqua, M., Gaetan, C., Mateu, J., and Porcu, E. (2012). Estimating space and space-time covariance functions for large data sets : a weighted composite likelihood approach. *Journal of the American Statistical Association*, 107(497) :268–280. [48](#), [50](#), [64](#)
- Castruccio, S., Huser, R., and Genton, M. (2015). High-order composite likelihood inference for max-stable distributions and processes. arXiv :1411.0086v3. [52](#)
- Cressie, N. (1993). *Statistics for Spatial Data : Wiley Series in Probability and Statistics*. Wiley-Interscience New York. [48](#)
- Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons. [41](#)
- Daley, D., Porcu, E., and Bevilacqua, M. (2014). Classes of compactly supported covariance functions for multivariate random fields. *Stochastic Environmental Research and Risk Assessment*, 29(4) :1249–1263. [42](#)
- De Iaco, S., Myers, D., Palma, M., and Posa, D. (2013). Using simultaneous diagonalization to identify a space-time linear coregionalization model. *Mathematical Geosciences*, 45(1) :69–86. [42](#)
- DLMF. NIST Digital Library of Mathematical Functions. <http://dlmf.nist.gov/>, Release 1.0.10 of 2015-08-07. [43](#)
- Gelfand, A. E. and Banerjee, S. (2010). Multivariate spatial process models. In Gelfand, A. E., Diggle, P., Fuentes, M., and Guttorp, P., editors, *Handbook of Spatial Statistics*, pages 495–515. Chapman & Hall/CRC. [41](#)
- Genton, M. G. and Kleiber, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statistical Science*, 30(2) :147–163. [41](#)
- Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97(458) :590–600. [42](#), [44](#), [46](#), [47](#), [65](#)
- Gneiting, T., Genton, M., and Guttorp, P. (2007). Geostatistical space-time models, stationarity, separability and full symmetry. In Finkenstädt, B., Held, L., and Isham, V., editors, *Statistical Methods for Spatio-Temporal Systems*, pages 151–175. Chapman & Hall/CRC. [42](#)

- Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491) :1167–1177. [42](#), [43](#), [44](#)
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477) :359–378. [54](#), [55](#)
- Gneiting, T. and Schlather, M. (2004). Stochastic models that separate fractal dimension and the Hurst effect. *SIAM Review*, 46(2) :269–282. [44](#)
- Goulard, M. and Voltz, M. (1992). Linear coregionalization model : Tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, 24(3) :269–286. [42](#)
- Gradshteyn, I. and Ryzhik, I. (2007). *Table of integrals, series, and products, Seventh edition*. Academic Press. [44](#), [65](#)
- Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press. [66](#)
- Li, B., Genton, M. G., and Sherman, M. (2008). Testing the covariance structure of multivariate random fields. *Biometrika*, 95(4) :813–829. [42](#)
- Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80(1) :221–239. [47](#)
- Matérn, B. (1986). *Spatial Variation*. Lecture Notes in Statistics, 36, Springer. [43](#)
- Porcu, E. and Schilling, R. L. (2011). From Schoenberg to Pick–Nevanlinna : Toward a complete picture of the variogram class. *Bernoulli*, 17(1) :441–455. [46](#)
- Porcu, E. and Zastavnyi, V. (2011). Characterization theorems for some classes of covariance functions associated to vector valued random fields. *Journal of Multivariate Analysis*, 102 :1293–1301. [42](#), [43](#), [45](#)
- Reisert, M. and Burkhardt, H. (2007). Learning equivariant functions with matrix valued kernels. *The Journal of Machine Learning Research*, 8 :385–408. [44](#)
- Schlather, M. (2010). Some covariance models based on normal scale mixtures. *Bernoulli*, 16(3) :780–797. [43](#), [46](#)
- Schoenberg, I. J. (1938). Metric spaces and completely monotone functions. *Annals of Mathematics*, 39(4) :811–841. [44](#)
- Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1) :5–42. [47](#)
- Wackernagel, H. (2003). *Multivariate geostatistics, Third Edition*. Springer, Berlin. [42](#)

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465) :250–261. [51](#)

Chapitre 4

Générateur stochastique de temps multisite

Sommaire

4.1 Introduction	72
4.2 Présentation de la zone d'étude et des données	73
4.3 Le modèle	76
4.3.1 Transformation des variables en résidus	76
4.3.2 Processus de précipitations	80
4.3.3 Structure et estimation du moment d'ordre 2 du champ gaussien $\mathbf{Y}$	83
4.4 Algorithme de simulation	90
4.4.1 Simulation non conditionnelle	91
4.4.2 Simulation conditionnelle	91
4.5 Validation du générateur stochastique de temps	93
4.5.1 Simulation non conditionnelle	93
4.5.2 Simulation conditionnelle	98
4.6 Conclusion et perspectives	104
4.7 Références	107

4.1 Introduction

Les SWGs multivariées [Richardson, 1981][Richardson and Wright, 1984][Semenov and Barrow, 1997] [Semenov et al., 1998][Rajagopalan and Lall, 1999][Flecher et al., 2010] ont d’abord été développés pour un seul site. Ces différents modèles reproduisent bien la structure temporelle de chaque variable ainsi que les corrélations entre celles-ci. En revanche si l’on génère des données en plusieurs sites d’une même région, ces modèles échoueront à reproduire correctement la structure spatiale existante puisque les simulations sont indépendantes d’un site à l’autre. Deux extensions [Wilks, 1999] [Khalili et al., 2009] du modèle de Richardson ont apporté une première réponse à cette problématique. Le processus stationnaire autorégressif multivarié d’ordre 1

$$Z_t = A Z_{t-1} + B \epsilon_t \quad (4.1)$$

est toujours au cœur du modèle pour simuler les résidus des variables (hors précipitation), où Z_t est un vecteur de dimension $p - 1$ (i.e. sans les précipitations puisque p est le nombre de variables), A, B des matrices de dimension $(p - 1) \times (p - 1)$ et ϵ_t un vecteur, de dimension $p - 1$, de nombres indépendants issus d’une loi normale $\mathcal{N}(0, 1)$. Ces matrices sont construites à partir des matrices de corrélation R_0 (lag-0) et R_1 (lag-1) avec

$$A = R_1 R_0^{-1} \quad \text{et} \quad B B^\top = R_0 - R_1 R_0^{-1} R_1. \quad (4.2)$$

Il est possible d’étendre ce cadre au cas spatial en passant de $p - 1$ à $(p - 1) \times n_s$, avec n_s le nombre de sites. Cependant l’accroissement du nombre de sites entraîne des difficultés pratiques, notamment pour garantir que la matrice $B B^\top$ soit semi-définie positive. Le processus de précipitation est basé à chaque fois sur un modèle markovien mais l’extension spatiale, au travers des probabilités de transition en chacun des sites, diffère entre les deux modèles : tirage dans une loi uniforme $[0, 1]$ corrélée spatialement [Khalili et al., 2007] et utilisation d’un vecteur gaussien corrélé spatialement [Wilks, 1998]. Les distributions de l’intensité des événements pluvieux s’appuient sur des lois exponentielles (simple ou double) dans les deux cas.

Un des avantages du modèle que nous présentons dans ce chapitre est sa simplicité. En effet, la structure spatio-temporelle de toutes les variables, y compris les précipitations, est guidée par un même et unique champ gaussien multivarié $\mathbf{Y}$. Ceci est rendu possible grâce à deux points importants. Le premier est le développement de nouvelles fonctions de covariance croisée permettant de prendre en compte la complexité des liens entre les variables, l’espace et le temps [Genton and Kleiber, 2014] [Bourotte et al., 2016]. En parallèle, des méthodes d’inférence adaptées aux grands jeux de données ont été élaborées comme l’approche par vraisemblance composite pour les champs gaussiens, qui sera utilisée dans ce travail en adaptant les équations de vraisemblance.

Le second point est la modélisation du processus de précipitations à l’aide d’une variable gaussienne. En effet, la pluie est une variable climatique très particulière puisque

c'est la combinaison d'une loi discrète pour modéliser l'occurrence ou non d'événements pluvieux, et d'une loi continue pour les cumuls de précipitations. Au premier abord, une variable gaussienne ne semble pas du tout la candidate adaptée à ce problème. Pourtant l'utilisation d'un champ gaussien latent permet à la fois de modéliser les occurrences et les intensités des précipitations [Allcroft and Glasbey, 2003] [Allard and Bourotte, 2014] [Baxevani and Lennatsson, 2015]. Les valeurs du champ gaussien en dessous d'un certain seuil correspondent à une absence de pluie tandis que les valeurs au dessus sont transformées en précipitation à l'aide d'une fonction adéquate, appelée fonction d'anamorphose [Chilès and Delfiner, 1999]. Cette idée, proposée par Guillot and Lebel [1999], présente deux avantages majeurs. Le premier est de permettre l'utilisation de différentes distributions marginales pour les intensités de pluie (loi Gamma, mélange GPD/Gamma, fonction polynomiale, ...). Deuxièmement, le champ latent est entièrement caractérisé par ses deux premiers moments, comme tout champ gaussien. Dans notre cas, le champ gaussien latent est la p -ième composante du champ gaussien multivarié $\mathbf{Y}$. Une alternative intéressante est de modéliser les seuils et les intensités par deux processus gaussiens indépendants [Barancourt et al., 1992]. Cette hypothèse a été reprise par Leblois and Creutin [2013] pour développer le générateur SAMPO, permettant la prise en compte de phénomène d'advection pour les précipitations. Cependant, nous voulions corrélérer les occurrences et les intensités de précipitations aux autres variables météorologiques, c'est pourquoi nous avons retenu la modélisation à l'aide d'un seul champ gaussien.

La zone d'étude de ce travail est la région méditerranéenne française et plus particulièrement les terres en bordure du Golfe du Lion. La répartition des stations et les particularités climatiques sont présentées dans la section 2. Le modèle global et l'estimation par maximum de vraisemblance (complète et par paires pondérées) sont décrits précisément dans la section 3. L'algorithme de simulation de données est détaillé pour la simulation conditionnelle et non conditionnelle. La section 5 offre un certain nombre de statistiques et de graphiques permettant de juger la capacité du générateur stochastique à reproduire différentes variables météorologiques en plusieurs sites d'une région pour le mois de janvier (les résultats concernant le mois de juillet sont ajoutés en annexe).

4.2 Présentation de la zone d'étude et des données

Le climat méditerranéen (en France) est propre au pourtour méditerranéen et au littoral corse. Les régions sont sous l'influence de la mer Méditerranée, chaude, et protégées des masses d'air venues de l'Atlantique ou du Nord par le relief : les contreforts orientaux des Pyrénées, la Montagne Noire, l'Espinouse, le rebord des causses du Larzac, les Cévennes, le sud des monts du Vivarais, les contreforts méridionaux des Alpes, du mont Ventoux à la frontière italienne. L'ensoleillement est important ; les hivers sont doux et les



FIGURE 4.1 – Localisation des stations sur la zone d'étude. Les noms en vert correspondent aux stations INRA et ceux en bleu aux stations Météo-France.

étés chauds. Ces régions sont les terres de prédilection de vents violents comme le mistral, la tramontane et le libeccio. Le régime pluviométrique est très particulier : seulement 70 à 80 jours de pluies supérieures à 1 millimètre (mm) irrégulièrement répartis dans l'année. A des étés secs succèdent des automnes très arrosés (40% du total annuel en 3 mois). Ces précipitations orageuses peuvent quelque fois apporter en quelques heures 4 fois plus que la moyenne mensuelle en un lieu donné.

Dans cette étude, 14 des 18 stations retenues sont situées en zone à climat méditerranéen et les quatre autres en périphérie : deux à l'ouest proche de Toulouse, une plus au nord dans le couloir rhodanien et la dernière en altitude (715 mètres, toutes les autres sont à moins de 200 mètres au dessus du niveau de la mer) au niveau des Causses. Les variables météorologiques disponibles pour chacun des 18 sites sont les températures minimales et maximales (T_n & T_x), la vitesse moyenne du vent (V), le rayonnement solaire (R) et les précipitations (P). La figure 4.1 présente la répartition des 18 stations sur la zone d'étude. Les indices des 14 stations à climat méditerranéen sont : 1, 3, 4, 6, 7, 8, 10, 11, 12, 13, 14, 15, 17, 18.

La pluviométrie et l'ensoleillement sont deux caractéristiques qui discriminent le climat méditerranéen des autres climats rencontrés sur le territoire français. Sur la période étudiée, les stations de Millau (12), Auzeville-Tolosane (31) et Lavaur (81) présentent un nombre de jours secs moins important que les autres sites, tous supérieurs à 70% de jours secs pour le mois de janvier, comme le montre le tableau 4.1. Pour le mois de juillet, les occurrences de pluies amènent au même type de remarque avec une proportion de jours secs proche des 90% pour les stations proche de la mer Méditerranée, et un pourcentage proche des 80% pour les 3 stations précédemment citées, ainsi que pour les deux

Station	1	2	3	4	5	6	7	8	9
% jours secs	0.73	0.65	0.76	0.73	0.70	0.80	0.76	0.74	0.52
Station	10	11	12	13	14	15	16	17	18
% jours secs	0.76	0.78	0.73	0.75	0.75	0.78	0.54	0.78	0.77
Station	1	2	3	4	5	6	7	8	9
% jours secs	0.92	0.79	0.92	0.82	0.79	0.90	0.88	0.91	0.79
Station	10	11	12	13	14	15	16	17	18
% jours secs	0.91	0.91	0.89	0.90	0.91	0.90	0.78	0.90	0.89

TABLE 4.1 – Proportion de jours secs pour chacune des stations sur les 11 mois de janvier (partie haute) et juillet (partie basse) (2003-2013).

stations drômoises. Cependant, la différence d'occurrence est un peu moins marquée en juillet qu'en janvier. La densité du rayonnement solaire est nettement bimodale pour les mois de janvier sur la zone étudiée avec un premier mode aux alentours de 300 J.cm^{-2} et un second vers 700 J.cm^{-2} . Les 14 stations à climat méditerranéen ont leur mode le plus important à 700 J.cm^{-2} tandis que les 4 stations périphériques ont le leur à 300 J.cm^{-2} . Cette différence est illustrée par la figure 4.2 avec une comparaison des densités pour les stations de Toulouse (Auzeville-Tolosane) et d'Avignon. La première ligne correspond au mois de janvier et la seconde à celui de juillet. Pour ce dernier, les deux densités sont unimodales (mode entre 2500 et 3000 J.cm^{-2}) avec une certaine asymétrie. La queue des densités est beaucoup plus lourde pour les faibles valeurs dans le cas de la station toulousaine par rapport à celle d'Avignon, ou des autres stations méditerranéennes. Au vu de ces résultats, les 14 stations à climat méditerranéen forment une zone homogène.

Dans ce chapitre, nous nous sommes concentrés sur les mois de janvier (pour des raisons pratiques liées à la rapidité d'exécution) mais le modèle proposé peut se décliner pour tous les mois de l'année ou les quatre saisons.

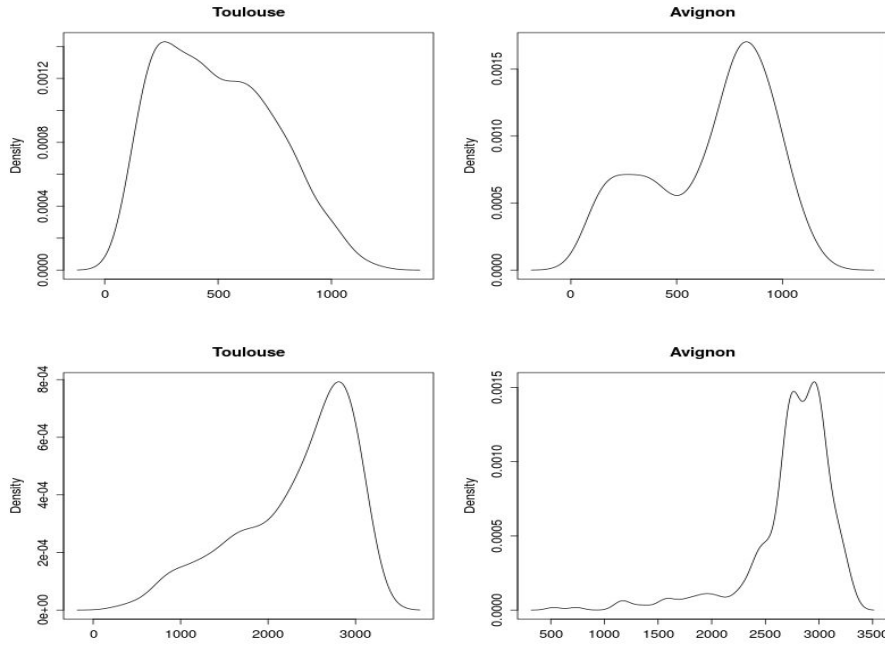


FIGURE 4.2 – Densité du rayonnement solaire pour les mois de janvier (ligne du haut) et juillet (ligne du bas) à Toulouse et Avignon (2003-2013).

4.3 Le modèle

La construction du générateur stochastique s'articule autour d'un champ gaussien stationnaire $\mathbf{Y}$ à p composantes, d'espérance nulle et de variance unité. Plusieurs transformations sont appliquées aux variables météorologiques $\mathbf{Z}$ pour obtenir le champ $\mathbf{Y}$. Les $p - 1$ premières composantes de $\mathbf{Y}$ sont utilisées pour modéliser le processus des résidus des variables météorologiques à l'exception des précipitations tandis que la dernière composante définit le processus gaussien latent qui, une fois transformé par la fonction d'anamorphose, produira les pluies ou leur absence. Le modèle génère des données journalières. Dans la suite, l'indice de temps t sera donc un entier naturel strictement positif correspondant à un jour et t_c sera sa position calendaire. La saisonnalité des variables météorologiques est prise en compte de manière classique au travers des fonctions de transformation du champ aléatoire $\mathbf{Y}$ et seront détaillées dans les différentes étapes de construction. La première sous section exposera la modification des $p - 1$ variables météorologiques en résidus, et les précipitations seront traitées dans la sous section suivante. Ensuite la caractérisation du champ $\mathbf{Y}$ par sa fonction de covariance croisée sera présentée, et nous finirons par un résumé du fonctionnement global du modèle.

4.3.1 Transformation des variables en résidus

La transformation des variables météorologiques en résidus se fait en plusieurs étapes (dans cette sous-section, nous n'abordons pas les précipitations). Une manière simple de traiter ce type de données est de considérer les données stationnaires sur le mois d'étude

et de procéder à une normalisation. Cette approche est tout à fait valide sur des périodes relativement courtes mais devient plus discutable lorsque l'on travaille sur une saison entière pour des variables comme les températures. Puisque notre générateur est censé fonctionner pour un mois comme pour une saison donnée, nous avons fait le choix de modéliser les tendances saisonnières à l'aide de fonctions sinusoïdales, proche de l'approche par splines proposée dans Richardson [1981] et Flecher et al. [2010].

Cependant avant de procéder à cette étape, nous commençons par corriger l'asymétrie de certaines distributions à l'aide de fonctions de transformation pour obtenir des marginales gaussiennes. Il est reconnu que les températures sont des variables météorologiques relativement symétriques contrairement aux vitesses de vent. Pour cette dernière variable, nous nous sommes appuyés sur la transformation Box-Cox [Box and Cox, 1964] :

$$\varsigma_0(x) = \log x, \quad (4.3)$$

$$\varsigma_\lambda(x) = \frac{x^\lambda - 1}{\lambda}, \quad \text{si } \lambda > 0, \quad (4.4)$$

avec $x > 0$, utilisée également dans Brown et al. [1984]. L'estimation du paramètre λ se fait par vraisemblance profilée à l'aide de la fonction *boxcox* du package MASS (sous R). Nous avons décidé de prendre une valeur commune de λ pour les 18 sites de la zone d'étude afin de préserver la structure spatiale des données de vent [Haslett and Raftery, 1989]. Ainsi le paramètre λ est régional, et nous avons choisi la moyenne des 18 estimations comme valeur commune [Bessac et al., 2015]. La moyenne estimée à partir des 18 valeurs de λ reste proche de zéro pour les différents mois de l'année et sur l'ensemble de l'année. Les vitesses de vent ont donc subi une transformation log. Le champ résultant est considéré comme gaussien, et une modélisation saisonnière de la moyenne et de la variance peut lui être appliquée [Rychlik, 2015]. La densité du rayonnement solaire présente aussi des asymétries mais, comme l'on peut le voir sur la figure 4.2, ces asymétries ne sont pas forcément les mêmes d'un mois à l'autre voire d'une station à l'autre. Dans ce chapitre, nous avons décidé de ne pas appliquer de transformation sur les marginales pour le rayonnement solaire. Dans la suite, on notera $g(Z_i)$, $i = 1, \dots, p-1$, où g est la fonction identité pour $i = 1, 2, 4$ et la fonction log pour $i = 3$.

Ensuite, il faut retirer la tendance saisonnière aux données transformées. Pour se faire, nous soustrayons une moyenne saisonnière $\mu_{i,\alpha}$ aux données transformées puis nous divisons par un écart-type saisonnier $\sigma_{i,\alpha}$ pour obtenir un premier champ intermédiaire Y' :

$$Y'_i(\mathbf{s}_\alpha, t) = \frac{g[Z_i(\mathbf{s}_\alpha, t)] - \mu_{i,\alpha}(t_c)}{\sigma_{i,\alpha}(t_c)}, \quad t_c = 1, \dots, 365, \quad t \in \mathbb{N}^*, \quad i = 1, \dots, p-1, \quad \alpha = 1, \dots, n_S, \quad \mathbf{s}_\alpha \in \mathbb{R}^d, \quad (4.5)$$

où n_S est le nombre de sites et $\mathbf{s}_\alpha$ les coordonnées des sites. Les deux fonctions $\mu_{i,\alpha}$ et $\sigma_{i,\alpha}$

sont modélisées par une fonction sinusoïdale f :

$$f_{i,\alpha}(t) = \bar{f}_{i,\alpha} + C_{i,\alpha}^f \cos \left\{ \frac{2\pi}{365} (t_c - T_{i,\alpha}^f) \right\}, \quad t_c = 1, \dots, 365, \quad t \in \mathbb{N}^*, \quad i = 1, \dots, p-1, \quad \alpha = 1, \dots, n_s, \quad (4.6)$$

avec $f = \{\mu, \sigma\}$. La valeur $\bar{f}_{i,\alpha}$ correspond à la moyenne des moyennes ou écarts-types des jours calendaires d'une variable en un site. Les paramètres $C_{i,\alpha}$ et $T_{i,\alpha}$ sont les coefficients de la fonction sinusoïdale correspondant respectivement à la valeur de crête et à sa position. Pour chacune des quatre variables, la moyenne et l'écart-type saisonniers ont été calculés deux fois en tous sites pour les jours pluvieux (wet) et les jours secs (dry) comme dans [Richardson \[1981\]](#) et [Wilks \[1999\]](#). La figure 4.3 illustre cette approche pour les températures minimales et le rayonnement solaire à Avignon de 2003 à 2013. Pour cette dernière variable, la ligne du bas de la figure 4.3 expose clairement l'intérêt de conditionner les paramètres saisonniers par l'occurrence de précipitation, ce qui devrait permettre au modèle de mieux rendre compte de la bimodalité de cette variable. Les quatre - deux pour les jours secs et deux pour les jours pluvieux - paramètres $C_{i,\alpha}$ et $T_{i,\alpha}$ sont estimés par la méthode des moindres carrés pour chacun des sites et chacune des variables hors précipitations. Cette méthode offre une très bonne adéquation entre les fonctions f et les données mais présente le défaut d'induire un nombre important de paramètres puisque les quatre paramètres $C_{i,\alpha}$ et $T_{i,\alpha}$ sont inférés en chacun des 18 sites pour les 4 variables (températures minimales et maximales, vitesse moyenne du vent, rayonnement solaire).

Il peut être intéressant d'obtenir ces paramètres pour un nouveau site sans données. Le krigeage peut fournir une interpolation spatiale de chacun de ces paramètres. Une solution alternative est d'exprimer les trois quantités $\bar{f}$, C et T en fonction de covariables spatiales comme la latitude et la longitude pour $d = 2$ ou la latitude, la longitude et l'altitude pour $d = 3$. Cette approche permet d'interpoler les valeurs de $\bar{f}$, C et T pour des sites sans données à partir des covariables et donc de calculer une moyenne et un écart-type saisonniers. Nous avons privilégié cette approche au début mais elle n'était pas applicable aux données de vitesse moyenne de vent de notre jeu de données. En effet, les vitesses ont été mesurées à différentes hauteurs (2 mètres pour les stations INRA, 10 mètres pour les stations Météo-France) ce qui faussait complètement l'estimation des paramètres. La première méthode permet de traiter indépendamment chaque station et donc de s'affranchir de cette hétérogénéité des protocoles de mesure.

Enfin, nous avons procédé à une dernière étape, même si celle-ci n'est pas indispensable. Les valeurs $Y'_i, i = 1, \dots, p-1$ ne sont pas tout à fait centrées ni réduites, que l'on travaille sur un mois donné ou une saison donnée. Or les équations de vraisemblance, présentées plus loin, sont beaucoup plus simples à écrire avec des processus centrés normés (la composante p pour la pluie l'est par construction). Donc pour chacune des variables en chacun des sites, une normalisation a été effectuée pour obtenir le champ $\mathbf{Y}$. Une future version du générateur intégrera directement cette normalisation dans la vraisemblance pour supprimer cette étape.

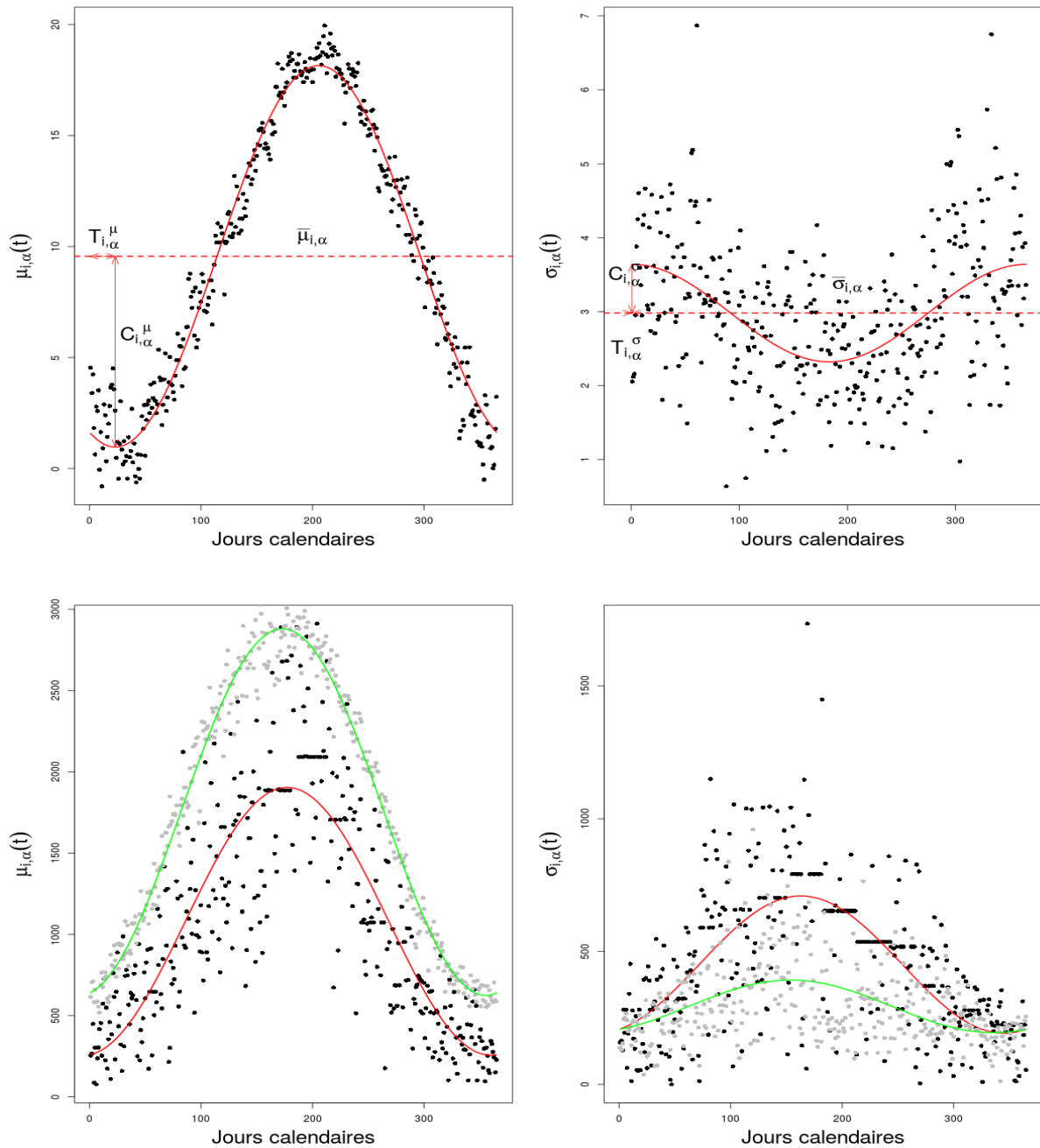


FIGURE 4.3 – Nuage de points : moyenne (à gauche) et écart-type (à droite) pour chaque jour calendaire des températures minimales des jours secs (ligne du haut) et du rayonnement solaire (ligne du bas) à Avignon de 2003 à 2013. En haut. Ligne rouge pointillée : moyenne du nuage de points. Courbe rouge : fonction sinusoïdale avec les paramètres C et T estimés. En bas. En gris les jours secs avec la fonction sinusoïdale correspondante en verte, et en noir les jours humides avec la fonction sinusoïdale correspondante en rouge.

4.3.2 Processus de précipitations

Les précipitations nécessitent une transformation différente du champ gaussien par rapport aux autres variables. Les données de pluie ne sont pas du tout gaussiennes et sont caractérisées par une masse importante de données égales à zéro, dépendant du pas de temps considéré (ici journalier), et des données strictement positives pour le reste. Nous suivons la même méthodologie que dans [Allard and Bourotte \[2014\]](#) ce qui implique que les précipitations sont pilotées par un processus gaussien latent centré normé auto-corrélé qui correspond ici à la p -ième composante du champ aléatoire $\mathbf{Y}$. Pour tout $(\mathbf{s}, t) \in \mathbb{R}^d \times \mathbb{R}$, la transformation des valeurs de $Y_p(\mathbf{s}, t)$ s'opère à l'aide d'une fonction de la forme $\Psi(y)I(y > y_0)$ où I est la fonction indicatrice et Ψ une fonction strictement croissante sur $]y_0, +\infty[$. La fonction de transformation Ψ est

$$\Psi_\alpha(y) = z_m + b_\alpha \left\{ e^{a_\alpha(y - y_{0,\alpha})^{c_\alpha}} - 1 \right\}, \quad y > y_{0,\alpha}, \quad \alpha = 1, \dots, n_S,$$

où z_m est la valeur minimale de précipitations correspondant à la précision de mesure des jauges pluviométriques. Dans la pratique les précipitations relevées sont des multiples de z_m . Le paramètre $y_{0,\alpha}$ est le seuil qui détermine directement l'occurrence ou non de précipitations au site $\mathbf{s}_\alpha$ ce qui se traduit par la relation suivante :

$$\mathbb{P}(\text{absence de pluie au site } \mathbf{s}_\alpha) = \mathbb{P}(Y_p(\mathbf{s}_\alpha, t) \leq y_{0,\alpha}) = \Phi(y_{0,\alpha}), \quad \alpha = 1, \dots, n_S, \quad \mathbf{s}_\alpha \in \mathbb{R}^d, \quad t \in \mathbb{N}^*, \quad (4.7)$$

où Φ est la fonction de répartition d'une loi normale $\mathcal{N}(0, 1)$. Ainsi le vecteur gaussien Y_p peut être vu comme un champ de potentiel : tant qu'il reste en dessous d'un seuil il ne pleut pas et quand il le dépasse, la valeur de potentiel est transformée en pluie par Ψ . La figure 4.4 permet de visualiser ce concept.

La capacité de Ψ à ajuster correctement la distribution marginale des intensités de précipitations journalières a été étudiée dans [Allard and Bourotte \[2014\]](#) sur des sites situés dans la Beauce, au sud-ouest de Paris, où le régime de pluies est très différent du régime méditerranéen, caractérisé par de fortes intensités de précipitations et des occurrences plus faibles. Cependant, comme nous allons l'illustrer plus loin, cette fonction, grâce aux trois paramètres a, b, c , est assez flexible pour ajuster aussi ce type de régime et reproduire des phénomènes de pluie extrêmes. En considérant à chaque fois les deux autres paramètres fixés, une interprétation du troisième paramètre peut être faite. Le paramètre a augmente l'étendue de la densité sans en modifier l'allure ; le paramètre b influe sur la forme du corps de la densité tandis que le paramètre c influe sur la queue.

L'indexation par α des paramètres $a_\alpha, b_\alpha, c_\alpha$ et $y_{0,\alpha}$ signifie que ces paramètres marginaux dépendent de la localisation. Pour rester cohérent avec le traitement des autres variables climatiques, nous avons décidé d'estimer ces paramètres pour chaque site $\mathbf{s}_\alpha$. Dans ce cas, la gestion de la saisonnalité des précipitations est effectuée de manière simple et relativement classique en se concentrant sur un mois ou une saison et en considé-

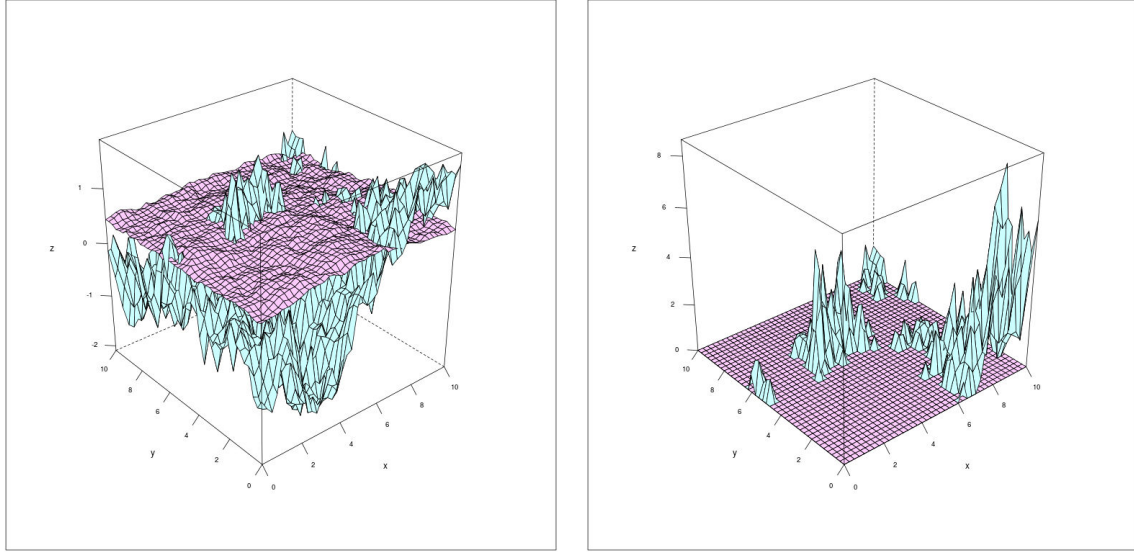


FIGURE 4.4 – A gauche : en bleu, champ gaussien d’espérance nulle, de variance unité et de fonction d’auto-corrélation $\exp(-\|\mathbf{h}\|/5)$, $\mathbf{h} \in \mathbb{R}^2$. En violet, surface qui définit la valeur du seuil en chaque point de l’espace. A droite : montants des précipitations suite à la transformation du champ gaussien. Les valeurs du champ gaussien inférieures à la surface seuil (surface rose à gauche) donnent des précipitations nulles (surface rose à droite). Les autres valeurs sont transformées en précipitations non nulles par la fonction Ψ avec $a = 5$, $b = 0.03$, $c = 0.4$.

rant les années comme des répétitions indépendantes. Cette méthode présente l’avantage d’offrir une très bonne adéquation avec les distributions marginales d’intensité de précipitations pour chaque site, au prix d’un nombre important de paramètres.

Ici aussi, les paramètres a, b, c et y_0 peuvent être interpolés en de nouveaux sites à l’aide d’une méthode de krigeage. Par exemple, la surface rose sur la figure gauche de 4.4 pourrait être une surface d’interpolation. Comme pour les autres variables, une alternative basée sur des fonctions sinusoïdales et des covariables spatiales pourraient être développée pour modéliser $a(\mathbf{s}, t)$, $b(\mathbf{s}, t)$, $c(\mathbf{s}, t)$ et $y_0(\mathbf{s}, t)$. C’est ce qui est proposé dans [Baxevani and Lennatsson, 2015] pour les paramètres de la loi Gamma et de la GPD ainsi que pour les valeurs de seuil (la moyenne du processus latent dans ce cas). Mais ce schéma nécessite de posséder pour chaque jour calendaire assez d’événements pluvieux pour procéder à l’inférence puisque seuls ces derniers sont utilisés dans la procédure d’estimation. Dans une région comme le sud-est de la France, même avec plusieurs dizaines d’années de données, ce n’est pas toujours possible pour certaines périodes sèches de l’année.

L’estimation des paramètres $a_\alpha, b_\alpha, c_\alpha$ et $y_{0,\alpha}$ se fait en deux étapes pour chacun des sites indépendamment des autres. Premièrement, l’estimation de la valeur seuil $y_{0,\alpha}$ se fait simplement en inversant l’équation 4.7 : $\hat{y}_{0,\alpha} = \Phi^{-1}(p_\alpha)$, $\alpha = 1, \dots, n_S$, où p_α est la proportion de zéros dans la série pluviométrique ce qui correspond à la probabilité d’absence de pluie au site concerné α . La seconde étape est une procédure d’estimation par maximum de vraisemblance. Le cadre est très classique, seules sont prises en compte les

Station	1	2	3	4	5	6	7	8	9
Paramètre a	5.52	6.08	5.90	6.65	5.38	6.93	6.50	6.68	5.54
Paramètre $b (\times 10^{-2})$	3.03	1.12	1.89	1.01	2.20	1.04	1.15	0.98	1.32
Paramètre $c (\times 10^{-1})$	4.46	3.36	3.88	3.34	4.46	2.63	3.57	2.99	4.07
Station	10	11	12	13	14	15	16	17	18
Paramètre a	5.84	6.94	6.39	5.50	5.72	6.21	5.56	7.08	6.97
Paramètre $b (\times 10^{-2})$	2.07	1.04	1.27	2.58	2.91	2.46	1.37	0.82	0.78
Paramètre $c (\times 10^{-1})$	4.57	3.17	4.17	4.95	5.50	4.77	3.55	3.23	3.12

TABLE 4.2 – Estimation par maximum de vraisemblance des paramètres a, b, c de la fonction Ψ pour chacune des stations sur les 11 mois de janvier (2003-2013).

valeurs de précipitation, i.e. les $Z_p(\mathbf{s}_\alpha, t) > 0$; de plus ces valeurs sont considérées comme indépendantes. La log-vraisemblance s'écrit alors :

$$l(\theta_\alpha) = - \sum_{Z_p(\mathbf{s}_\alpha, t) > 0} \left\{ \frac{[\Psi_\alpha^{-1}(Z_p(\mathbf{s}_\alpha, t))]^2}{2} + \frac{1}{c_\alpha} \log a_\alpha + \log c_\alpha + \log(Z_p(\mathbf{s}_\alpha, t) - z_m^* + b_\alpha) \right. \\ \left. + (1 - \frac{1}{c_\alpha}) \log \left[\log \left(\frac{Z_p(\mathbf{s}_\alpha, t) - z_m^*}{b_\alpha} + 1 \right) \right] \right\}, \quad \alpha \text{ fixé}, t \in \mathbb{N}^*, \quad (4.8)$$

avec $\theta_\alpha = (a_\alpha, b_\alpha, c_\alpha)^\top$. Les termes de l'équation après le carré de Ψ_α^{-1} correspondent au log du jacobien. On peut remarquer que nous avons remplacé z_m par z_m^* puisque les valeurs minimales de précipitation sont égales à z_m . Sans cela, il résulterait une forme indéterminée dans la seconde ligne de l'équation 4.8. Pour nos données, la valeur de discrétisation est $z_m = 0.5$ mm et nous avons obtenu les meilleurs résultats avec $z_m^* = 0.375$ mm. La maximisation de $l(\theta_\alpha)$ fournit une estimation $\hat{\theta}_\alpha$ pour chacun des 18 sites. Le tableau 4.2 présente les résultats de la procédure d'estimation pour les paramètres a, b, c ; pour obtenir les estimations du paramètre y_0 , il suffit d'appliquer la transformation Φ^{-1} aux valeurs du tableau 4.1. Les quantile-quantile plots ou *qq-plots* sont très utiles pour juger l'adéquation entre les données de pluie et le modèle de précipitations avec les paramètres estimés. Nous avons simulé 100 séries de précipitations pour chacun des sites pour les mois de janvier. La figure 4.5 présentent ces résultats. Le modèle de précipitation basé sur la fonction d'anamorphose ψ permet de simuler aussi bien les petites que les grandes valeurs de précipitations observées, quelque soit le mois ou la saison (voir en annexe pour le mois de juillet).

La valeur du paramètre c permet d'adapter la distribution de l'intensité des précipitations aux valeurs extrêmes. Pour les mois de janvier, les trois plus hautes valeurs estimées du paramètre c correspondent aux stations où l'on a enregistré

- deux événements pluvieux supérieurs à 40 mm et un à 90 mm (pour Roujan, station 13) ;
- trois événements pluvieux supérieurs à 40 mm et un à 100 mm (pour Alenya, station

14);

- trois événements pluvieux supérieurs à 40 mm et un à 120 mm (pour Perpignan, station 15);

tandis que les deux plus basses valeurs estimées correspondent aux stations Bellegarde (station 6) et Saint Gilles (station 8) où le plus fort événement pluvieux est d'environ 35 mm (tous les autres événements pluvieux sont inférieurs à 25 mm).

4.3.3 Structure et estimation du moment d'ordre 2 du champ gaussien $\mathbf{Y}$

La dernière étape de la construction du modèle consiste à définir la structure d'ordre 2 du champ gaussien $\mathbf{Y}$. Nous choisissons de travailler dans un cadre paramétrique et de nous appuyer sur les travaux du chapitre 3, ce qui nous amène à faire les deux hypothèses suivantes. Premièrement, la fonction de covariance croisée - le champ est multivarié - est stationnaire, ce qui ne semble pas trop restrictif ici puisque la zone d'étude présente une bonne homogénéité (cf section 4.2). Deuxièmement, la fonction de covariance croisée est isotropique. Cette hypothèse sera discutée plus loin.

A l'instar des données du chapitre 3, la décomposition du champ $\mathbf{Y}$ en somme de deux champs gaussiens indépendants permet de s'adapter au mieux aux données de la zone d'étude. Ainsi le premier champ gaussien $\mathbf{X}$ est un champ temporel multivarié c'est à dire qu'il se comporte de la même façon en tout point de la zone et peut être considéré comme un processus régional. Le second champ $\mathbf{W}$ intègre en plus la dimension spatiale, tel que :

$$Y_i(\mathbf{s}, t) = X_i(t) + W_i(\mathbf{s}, t), \quad i, j = 1, \dots, 5, \quad (\mathbf{s}, t) \in \mathbb{R}^2 \times \mathbb{R}. \quad (4.9)$$

Les trois champs $\mathbf{X}, \mathbf{W}, \mathbf{Y}$ sont bien évidemment d'espérance nulle. Rychlik [2015] utilise également une somme de quatre champs gaussiens indépendants pour modéliser les vitesses de vent.

Pour la modélisation de données météorologiques, il est souvent préférable d'utiliser des modèles de covariance non séparables, comme pour les vitesses moyennes de vent dans [Gneiting, 2002] ou pour les températures et le rayonnement solaire au chapitre précédent. Les résultats encourageants sur les données de l'Ouest de la France nous ont motivé à réutiliser les mêmes modèles de covariance croisée pour les données médi-

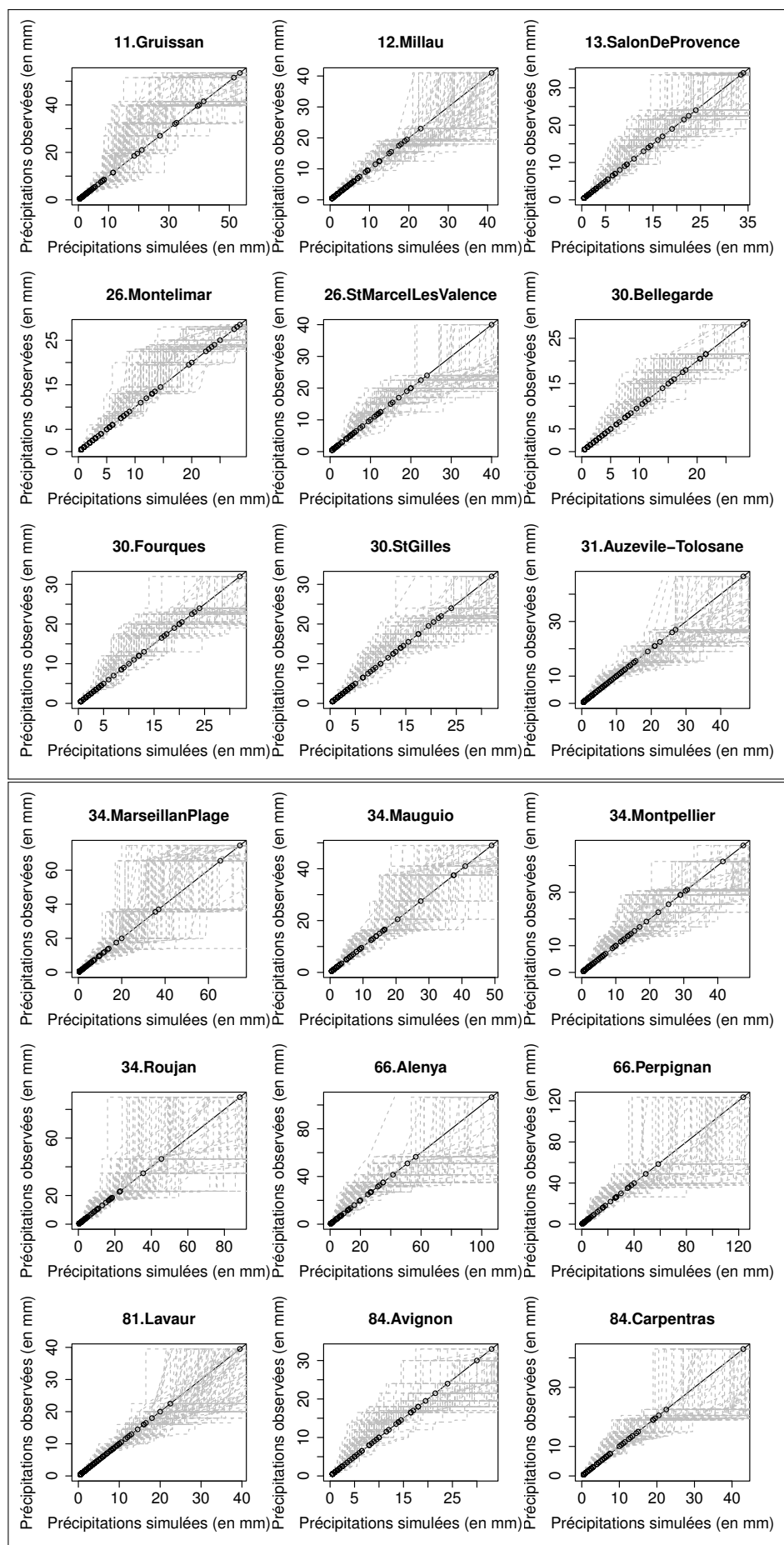


FIGURE 4.5 – QQ-plots : précipitations observées vs précipitations simulées (100 simulations) pour le mois de janvier. Les points sur la diagonale correspondent aux valeurs observées.

terraneennes, et qui s'écrivent :

$$\begin{aligned}
 C_{ij}^{(X)}(u) &= \sigma_i^{(X)} \sigma_j^{(X)} \beta_{ij}^{(X)} \left(\alpha^{(X)} |u|^{2a^{(X)}} + 1 \right)^{-1}, \\
 C_{ij}^{(W)}(\mathbf{h}, u) &= \sigma_i^{(W)} \sigma_j^{(W)} \beta_{ij}^{(W)} \left(\alpha^{(W)} |u|^{2a^{(W)}} + 1 \right)^{-1} \frac{\Gamma\{(v_i + v_j)/2\}}{\Gamma(v_i)^{1/2} \Gamma(v_j)^{1/2}} \frac{r_i^{v_i} r_j^{v_j}}{\{(r_i^2 + r_j^2)/2\}^{(v_i + v_j)/2}} \\
 &\quad \times \mathcal{M} \left(\frac{\mathbf{h}}{(\alpha^{(W)} |u|^{2a^{(W)}} + 1)^{b/2}} \middle| \sqrt{\frac{r_i^2 + r_j^2}{2}}, \frac{v_i + v_j}{2} \right), \quad i, j = 1, \dots, 5, \quad (\mathbf{h}, u) \in \mathbb{R}^2 \times \mathbb{R}.
 \end{aligned} \tag{4.10}$$

De même la contrainte $\left[\sigma_i^{(X)} \right]^2 + \left[\sigma_i^{(W)} \right]^2 = 1$, avec $i = 1, \dots, 5$, a été renouvelée ici pour deux raisons. La première est que, par construction, le vecteur gaussien latent pour les précipitations est de variance unité. Donc pour faciliter l'écriture des différentes vraisemblances - qui sont présentées un peu plus loin - les quatre autres variables ont vu aussi leurs variances être fixées à un. Deuxièmement, le nombre de paramètres du modèle est très élevé et donc il est toujours intéressant de pouvoir le réduire pour économiser en temps de calcul et faciliter l'étape d'optimisation. Dans la suite,

$$C_{ij}(\mathbf{h}, u) = C_{ij}^{(X)}(u) + C_{ij}^{(W)}(\mathbf{h}, u), \quad i, j = 1, \dots, 5, \quad (\mathbf{h}, u) \in \mathbb{R}^2 \times \mathbb{R}, \tag{4.11}$$

fait référence à la fonction de covariance croisée de $\mathbf{Y}$.

Maintenant que la structure d'ordre 2 du champ gaussien multivarié $\mathbf{Y}$ est clairement définie, il reste à présenter la méthode d'estimation des différents paramètres θ des deux fonctions de covariance croisée. Les résultats du chapitre 3 ont démontré l'efficacité de l'estimation de fonctions de covariance croisée complexe par l'approche par vraisemblance marginale par paires dans le cadre de champs gaussiens. Nous allons de nouveau nous appuyer sur cette méthode pour procéder à l'estimation des paramètres. Cependant, l'écriture de la vraisemblance marginale d'une paire va être un peu plus compliquée ici à cause des précipitations.

En effet, on observe le vecteur $\mathbf{Z} = (Z_1(\mathbf{s}_1, t_1), \dots, Z_5(\mathbf{s}_1, t_1), \dots, Z_5(\mathbf{s}_n, t_n))^T$, qui correspond aux données météorologiques des cinq variables journalières suivantes : température minimale, température maximale, vitesse moyenne du vent, rayonnement solaire et précipitations. Une fois les données transformées, nous obtenons le vecteur $(Y_1(\mathbf{s}_1, t_1), \dots, Y_4(\mathbf{s}_1, t_1), Z_5(\mathbf{s}_1, t_1), \dots, Z_5(\mathbf{s}_n, t_n))^T$ avec $Y_i(\mathbf{s}_\alpha, t_\alpha)$, $i = 1, \dots, 4$, $\alpha = 1, \dots, n$, les résidus centrés normés. Les précipitations se divisent en deux cas : soit $Z_5(\mathbf{s}_\alpha, t_\alpha)$, $\alpha = 1, \dots, n$, est strictement positif et l'on peut alors le transformer en un résidu $Y_5(\mathbf{s}_\alpha, t_\alpha)$ à l'aide de la transformation inverse ψ^{-1} ; soit $Z_5(\mathbf{s}_\alpha, t_\alpha)$, $\alpha = 1, \dots, n$, est nul et la valeur du résidu est censurée. Nous sommes donc amenés à considérer six cas distincts de log-vraisemblance $l(i, j, \mathbf{s}_\alpha, \mathbf{s}_\beta, t_\alpha, t_\beta; \theta)$, appelée brique élémentaire au chapitre 2, en fonction de la paire choisie ($\alpha, \beta = 1, \dots, n$ et $i, j = 1, \dots, 5$). Pour les cas avec des valeurs censurées, nous suivons les équations proposées dans [Allcroft and Glasbey \[2003\]](#). Ainsi,

— la paire est du type $(Y_i(\mathbf{s}_\alpha, t_\alpha), Y_j(\mathbf{s}_\beta, t_\beta))$ avec $i, j = 1, \dots, p-1$:

$$l(i, j, s_\alpha, s_\beta, t_\alpha, t_\beta; \theta) = -\frac{1}{2} \left\{ \log \Delta_{ij, \alpha\beta} + \frac{Y_i(\mathbf{s}_\alpha, t_\alpha)^2 - 2C_{ij}(\mathbf{h}, u)Y_i(\mathbf{s}_\alpha, t_\alpha)Y_j(\mathbf{s}_\beta, t_\beta) + Y_j(\mathbf{s}_\beta, t_\beta)^2}{\Delta_{ij, \alpha\beta}} \right\} \quad (4.12)$$

— la paire est du type $(Y_i(\mathbf{s}_\alpha, t_\alpha), Z_p(\mathbf{s}_\beta, t_\beta))$ avec $i = 1, \dots, p-1$ et $Z_p(\mathbf{s}_\beta, t_\beta) = 0$:

$$l(i, p, s_\alpha, s_\beta, t_\alpha, t_\beta; \theta) = \log \Phi \left(\frac{y_{0,\beta} - C_{ip}(\mathbf{h}, u)Y_i(\mathbf{s}_\alpha, t_\alpha)}{\Delta_{ip, \alpha\beta}^{1/2}} \right) \quad (4.13)$$

— la paire est du type $(Y_i(\mathbf{s}_\alpha, t_\alpha), Z_p(\mathbf{s}_\beta, t_\beta))$ avec $i = 1, \dots, p-1$ et $Z_p(\mathbf{s}_\beta, t_\beta) > 0$:

$$Y_p(\mathbf{s}_\beta, t_\beta) = \psi_\beta^{-1}(Z_p(\mathbf{s}_\beta, t_\beta))$$

$$l(i, p, s_\alpha, s_\beta, t_\alpha, t_\beta; \theta) = -\frac{1}{2} \left\{ \log \Delta_{ip, \alpha\beta} + \frac{Y_i(\mathbf{s}_\alpha, t_\alpha)^2 - 2C_{ip}(\mathbf{h}, u)Y_i(\mathbf{s}_\alpha, t_\alpha)Y_p(\mathbf{s}_\beta, t_\beta) + Y_p(\mathbf{s}_\beta, t_\beta)^2}{\Delta_{ip, \alpha\beta}} \right\} \quad (4.14)$$

— la paire est du type $(Z_p(\mathbf{s}_\alpha, t_\alpha), Z_p(\mathbf{s}_\beta, t_\beta))$ avec $Z_p(\mathbf{s}_\alpha, t_\alpha) = 0$ et $Z_p(\mathbf{s}_\beta, t_\beta) = 0$:

$$l(p, p, s_\alpha, s_\beta, t_\alpha, t_\beta; \theta) = \log \Phi_2(y_{0,\alpha}, y_{0,\beta}; C_{pp}(\mathbf{h}, u)) \quad (4.15)$$

où Φ_2 est la fonction de répartition de la loi normale bivariee.

— la paire est du type $(Z_p(\mathbf{s}_\alpha, t_\alpha), Z_p(\mathbf{s}_\beta, t_\beta))$ avec $Z_p(\mathbf{s}_\alpha, t_\alpha) = 0$ et $Z_p(\mathbf{s}_\beta, t_\beta) > 0$:

$$Y_p(\mathbf{s}_\beta, t_\beta) = \psi_\beta^{-1}(Z_p(\mathbf{s}_\beta, t_\beta))$$

$$l(p, p, s_\alpha, s_\beta, t_\alpha, t_\beta; \theta) = \log \Phi \left(\frac{y_{0,\alpha} - C_{pp}(\mathbf{h}, u)Y_p(\mathbf{s}_\beta, t_\beta)}{\Delta_{pp, \alpha\beta}^{1/2}} \right) \quad (4.16)$$

— la paire est du type $(Z_p(\mathbf{s}_\alpha, t_\alpha), Z_p(\mathbf{s}_\beta, t_\beta))$ avec $Z_p(\mathbf{s}_\alpha, t_\alpha) > 0$ et $Z_p(\mathbf{s}_\beta, t_\beta) > 0$:

$$Y_p(\mathbf{s}_\alpha, t_\alpha) = \psi_\alpha^{-1}(Z_p(\mathbf{s}_\alpha, t_\alpha))$$

$$Y_p(\mathbf{s}_\beta, t_\beta) = \psi_\beta^{-1}(Z_p(\mathbf{s}_\beta, t_\beta))$$

$$l(p, p, s_\alpha, s_\beta, t_\alpha, t_\beta; \theta) = -\frac{1}{2} \left\{ \log \Delta_{pp, \alpha\beta} + \frac{Y_p(\mathbf{s}_\alpha, t_\alpha)^2 - 2C_{pp}(\mathbf{h}, u)Y_p(\mathbf{s}_\alpha, t_\alpha)Y_p(\mathbf{s}_\beta, t_\beta) + Y_p(\mathbf{s}_\beta, t_\beta)^2}{\Delta_{pp, \alpha\beta}} \right\} \quad (4.17)$$

où $\Delta_{ij, \alpha\beta} = 1 - C_{ij}(\mathbf{h}, u)^2$ avec $\mathbf{h} = \|\mathbf{s}_\alpha - \mathbf{s}_\beta\|$ et $u = |t_\alpha - t_\beta|$. Nous rappelons que le champ gaussien $\mathbf{Y}$ est de variance unité et donc que $C_{ii}(\mathbf{0}, 0) = 1$ pour tout $i = 1, \dots, 5$. L'écriture des équations est faite dans le cas général où $p \in \mathbb{N}^*$ avec les précipitations en dernière position (i.e. p -ième position).

Il ne reste plus qu'à sommer l'ensemble de ces briques élémentaires pour obtenir la vraisemblance marginale par paires pondérées $wpl(\theta)$ [Bevilacqua and Gaetan, 2014] :

$$wpl(\theta) = \sum_{(i, j, \alpha, \beta) \in \Lambda} l(i, j, \mathbf{s}_\alpha, \mathbf{s}_\beta, t_\alpha, t_\beta; \theta) w_{ij, \alpha\beta}, \quad (4.18)$$

$\sigma_{Tn}^{(X)}$	$\sigma_{Tx}^{(X)}$	$\sigma_V^{(X)}$	$\sigma_R^{(X)}$	$\sigma_P^{(X)}$	$\sigma_{Tn}^{(W)}$	$\sigma_{Tx}^{(W)}$	$\sigma_V^{(W)}$	$\sigma_R^{(W)}$	$\sigma_P^{(W)}$
0.669	0.846	0.241	0.000	0.621	0.744	0.533	0.970	1.000	0.783
$\beta_{Tn/Tx}^{(X)}$	$\beta_{TnV}^{(X)}$	$\beta_{TnR}^{(X)}$	$\beta_{TnP}^{(X)}$	$\beta_{TxV}^{(X)}$	$\beta_{TxR}^{(X)}$	$\beta_{TxP}^{(X)}$	$\beta_{VR}^{(X)}$	$\beta_{VP}^{(X)}$	$\beta_{RP}^{(X)}$
0.959	0.006	-0.017	-0.259	-0.214	-0.018	-0.271	0.121	0.061	0.134
$\beta_{TnTx}^{(W)}$	$\beta_{TnV}^{(W)}$	$\beta_{TnR}^{(W)}$	$\sigma_{TnP}^{(W)}$	$\sigma_{TxV}^{(W)}$	$\sigma_{TxR}^{(W)}$	$\beta_{TxP}^{(W)}$	$\beta_{VR}^{(W)}$	$\beta_{VP}^{(W)}$	$\beta_{RP}^{(W)}$
0.377	0.108	-0.346	0.059	-0.424	0.020	-0.068	0.190	0.055	-0.261
v_{Tn}	v_{Tx}	v_V	v_R	v_P	$1/r_{Tn}$	$1/r_{Tx}$	$1/r_V$	$1/r_R$	$1/r_P$
0.288	0.307	0.263	0.262	0.501	313	218	285	196	140
$\alpha^{(X)}$	$a^{(X)}$	$\alpha^{(W)}$	$a^{(W)}$	b					
0.231	0.641	1.215	0.999	0.600					

TABLE 4.3 – Estimation par maximum de vraisemblance par paires pondérées des paramètres de la fonction de covariance croisée pour le mois de janvier (2003-2013) avec $\mathbf{d} = (150 \text{ km}, 2 \text{ days})$. Les distances sont en kilomètres.

	σ_{TnTx}	σ_{TnV}	σ_{TnR}	σ_{TxV}	σ_{TxR}	σ_{VR}
Processus $\mathbf{X}$	0.54	0.01	0.00	-0.04	0.00	0.00
Processus $\mathbf{W}$	0.15	0.08	-0.25	-0.22	0.01	0.18
Processus $\mathbf{Y}$	0.69	0.09	-0.25	-0.26	0.01	0.18
Empirique	0.62	0.22	-0.18	-0.13	0.13	0.09

TABLE 4.4 – Coefficients de covariance croisée empiriques et estimés $\sigma_{ij} = \sigma_i \sigma_j \rho_{ij}$ des résidus (hors précipitation), avec $i, j \in \{Tn, Tx, V, R\}$.

où

$$\Lambda = \left\{ (i, j, \alpha, \beta) : \begin{array}{l} i = j = 1, \dots, p, \quad \alpha = 1, \dots, n-1, \quad \beta = \alpha + 1, \dots, n, \\ i = 1, \dots, p-1, \quad j = i + 1, \dots, p, \quad \alpha, \beta = 1, \dots, n. \end{array} \right.$$

Dans un soucis de simplicité, la pondération choisie ne dépend pas des variables mais uniquement du vecteur $(\mathbf{h}, u)$, comme au chapitre 3. La pondération $w_{ij, \alpha\beta}$ est binaire : elle vaut 1 si $\mathbf{h}$ est inférieur ou égal à 150 kilomètres et u à 2 jours, et 0 sinon. Les simulations du chapitre 3 ont démontré qu'il n'était pas nécessaire de prendre en compte un pourcentage important de paires pour obtenir des estimations efficaces. Ici, le pourcentage de paires utilisées est de 10.1% (13.5% au chapitre 3) ce qui permet un gain considérable en terme de temps de calcul tout en garantissant de bonnes performances d'estimation. Pour chacune des variables, la portée pratique, calculée empiriquement, est atteinte au moins à 90% avec la valeur de 150 kilomètres.

Les valeurs estimées de la fonction de covariance croisée 4.10 sont présentées dans le tableau 4.3. Le nombre de paramètres est de 45, et seuls 40 sont estimés du fait de

la contrainte (les variances somment à 1). Le processus temporel $\mathbf{X}$ apporte une faible contribution pour les variables V et R , comme le montrent les valeurs $\sigma_V^{(X)}$ et $\sigma_R^{(X)}$, contrairement aux températures maximales. La contribution est équilibrée pour les deux variables restantes (T_n et P). En conséquence, les valeurs des coefficients $\sigma_{ij}^{(X)}$, avec $i, j \in \{T_n, T_x, V, R\}$, sont quasiment toutes nulles (tableau 4.4) sauf pour $\sigma_{T_n T_x}^{(X)}$. Les différents paramètres de régularité estimés sont assez proches (hors précipitation), mais les paramètres de portée affichent de grandes différences selon les variables. Enfin, le paramètre b de séparabilité est de 0.6 avec $b \in [0, 1]$ et où $b = 0$ correspond au modèle séparable pour $C_{ij}^{(W)}$.

Les figures 4.6 et 4.7 permettent de visualiser la bonne adéquation entre le modèle de covariance 4.10 avec les paramètres du tableau 4.3 et les résidus (processus $\mathbf{Y}$) des variables T_n , T_x , V et R . Pour la figure 4.6, les covariances empiriques directes et croisées sont calculées en fixant le temps, tandis que pour la figure 4.7, les covariances sont calculées pour une distance fixe. L'adéquation est bonne pour les différentes variables, même si le rayonnement solaire (R) pose quelques difficultés. En effet, la décroissance temporelle du rayonnement est très forte puisque pour $u = 1$ jour, la covariance est déjà presque nulle (case R pour les deux figures), ce qui n'est pas le cas des autres variables. C'est une des limitations du modèle de covariance présenté dans cette thèse. Il n'autorise pas à chaque variable d'avoir sa propre dynamique temporelle. Le fait d'utiliser une somme de covariances 4.9 permet d'apporter une certaine flexibilité et de limiter ce problème.

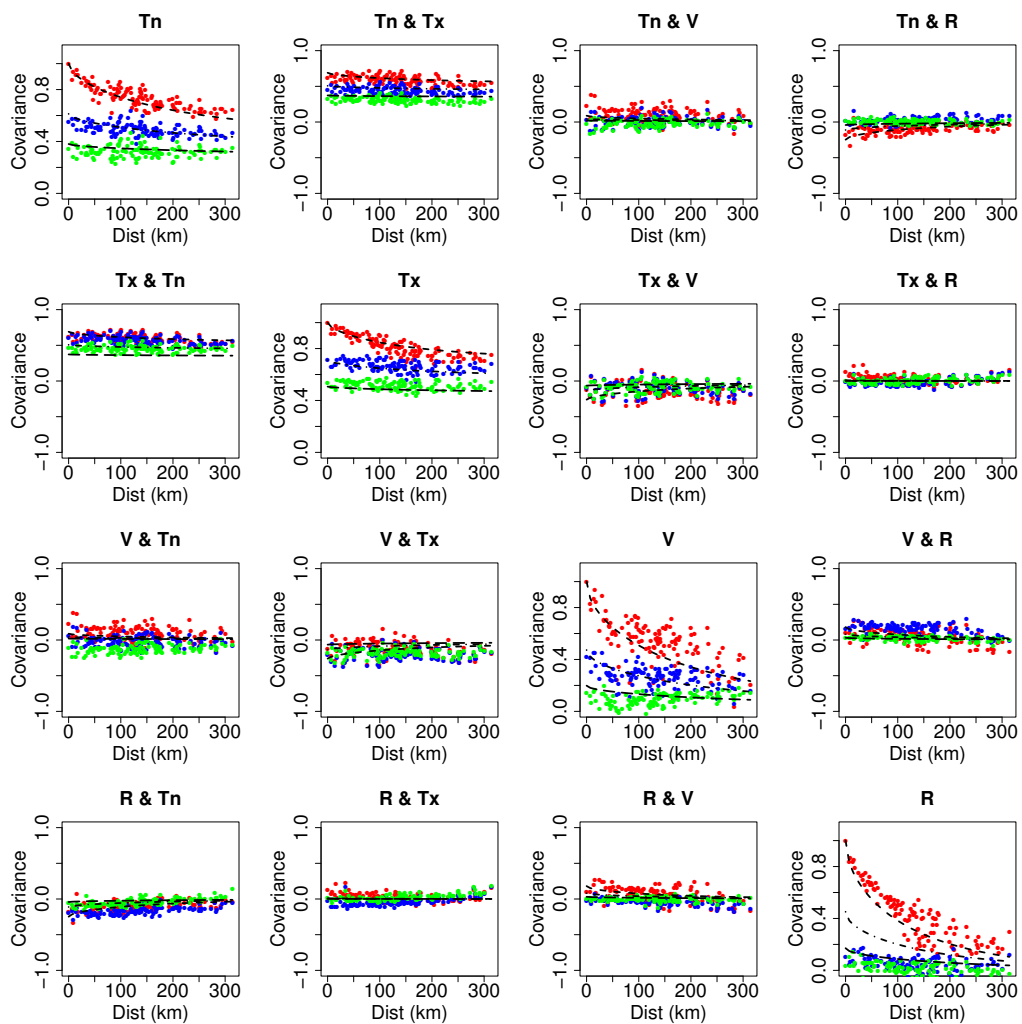


FIGURE 4.6 – Adéquation entre les covariances empiriques directes et croisées des résidus des variables (hors précipitation) et le modèle de covariance proposé 4.10 avec les paramètres estimés (tableau 4.3). Le décalage temporel u est fixé à 0 pour les nuages rouges, 1 pour les bleus et 2 pour les verts.

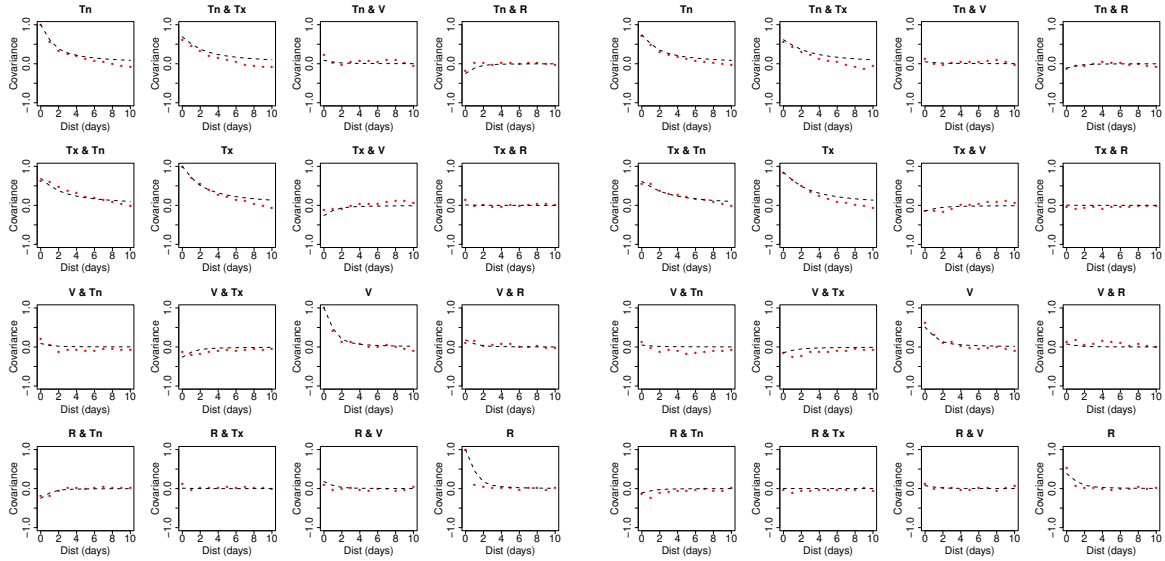


FIGURE 4.7 – Adéquation entre les covariances empiriques directes et croisées des résidus des variables (hors précipitation) et le modèle de covariance proposé 4.10 avec les paramètres estimés (tableau 4.3). A gauche : $\mathbf{h} = 0$. A droite : $\|\mathbf{h}\| = 100$ km.

4.4 Algorithme de simulation

Comme nous venons de le montrer, la construction de notre générateur stochastique de temps s'articule autour d'un champ gaussien multivarié. Outre sa simplicité, cette construction permet également d'effectuer des simulations conditionnelles, en plus de la génération classique de séries météorologiques (simulation non conditionnelle). Ces deux types de simulation répondent à des demandes différentes de la part de l'utilisateur. Pour illustrer ces propos, prenons l'exemple des mois de janvier (31 jours) sur notre zone d'étude.

La simulation non conditionnelle est le cœur de métier d'un SWG. Il est possible de simuler pour chacun des 18 sites les cinq variables d'intérêt sur une période de taille définie - petite ou grande. Les mois de janvier simulés sont indépendants les uns des autres. Il est également possible de simuler en d'autres points de la région du moment que l'on possède une méthode pour calculer les coefficients saisonniers en ces nouveaux points (surface d'interpolation paramétrique, krigeage ...).

La simulation conditionnelle permet de générer ces mêmes variables tout en prenant en considération les valeurs des observations en plusieurs points de la zone. Cela peut être très utile dans plusieurs situations. Par exemple, une des 18 stations subit une maintenance durant le mois de janvier 2014. Ces données manquantes peuvent être complétées par des données simulées conditionnellement aux valeurs observées aux 17 autres stations pendant le mois de janvier 2014. On peut procéder de la même façon avec un ou plusieurs sites autres que les 18 référencés.

4.4.1 Simulation non conditionnelle

La première étape de l'algorithme de simulation consiste à générer un vecteur auto-corrélé $\mathbf{Y} = (Y_1(\mathbf{s}_1, t_1), \dots, Y_p(\mathbf{s}_n, t_n))^T$ de résidus. Ici, n , nombre de points spatio-temporels, vaut 558 (31×18 , avec $n_s = 18$ le nombre de sites) et p , le nombre de variables, est égal à 5. Donc le vecteur $\mathbf{Y}$ est de longueur $np = 2790$, et la matrice Σ , matrice de variance-covariance de $\mathbf{Y}$, est de dimension $np \times np$. Le vecteur $\mathbf{Y}$ est obtenu en multipliant la matrice triangulaire inférieure L , issue de la décomposition de Cholesky de la matrice $\Sigma = LL^T$, avec le vecteur u de longueur np , composé d'éléments indépendants tirés dans une loi normale $\mathcal{N}(0, 1)$. A tout point spatio-temporel d'indice compris entre 1 et n correspond un indice de sites α , avec $\alpha = 1, \dots, n_s$, et de jours t , avec $t = 1, \dots, 31$. La seconde étape transforme le processus $\mathbf{Y}$ en variables météorologiques $\mathbf{Z}$ selon ce schéma :

$$Z_i(\mathbf{s}_\alpha, t) = \begin{cases} \mu_{i,\alpha}^W(t) + \sigma_{i,\alpha}^W(t) [m_{i,\alpha} + sd_{i,\alpha} Y_i(\mathbf{s}_\alpha, t)], & \text{si } i = 1, 2, 4 \text{ et } Y_5(\mathbf{s}_\alpha, t) > y_{0,\alpha}; \\ \mu_{i,\alpha}^D(t) + \sigma_{i,\alpha}^D(t) [m_{i,\alpha} + sd_{i,\alpha} Y_i(\mathbf{s}_\alpha, t)], & \text{si } i = 1, 2, 4 \text{ et } Y_5(\mathbf{s}_\alpha, t) \leq y_{0,\alpha}; \\ \exp \left\{ \mu_{i,\alpha}^W(t) + \sigma_{i,\alpha}^W(t) [m_{i,\alpha} + sd_{i,\alpha} Y_i(\mathbf{s}_\alpha, t)] \right\}, & \text{si } i = 3 \text{ et } Y_5(\mathbf{s}_\alpha, t) > y_{0,\alpha}; \\ \exp \left\{ \mu_{i,\alpha}^D(t) + \sigma_{i,\alpha}^D(t) [m_{i,\alpha} + sd_{i,\alpha} Y_i(\mathbf{s}_\alpha, t)] \right\}, & \text{si } i = 3 \text{ et } Y_5(\mathbf{s}_\alpha, t) \leq y_{0,\alpha}; \\ z_m + b_\alpha \{ \exp [a_\alpha (Y_i(\mathbf{s}_\alpha, t) - y_{0,\alpha})^{c_\alpha}] - 1 \}, & \text{si } i = 5 \text{ et } Y_5(\mathbf{s}_\alpha, t) > y_{0,\alpha}; \\ 0, & \text{si } i = 5 \text{ et } Y_5(\mathbf{s}_\alpha, t) \leq y_{0,\alpha}. \end{cases}$$

Nous rappelons que les termes μ et σ proviennent de la modélisation saisonnière (cf figure 4.3) tandis que les termes m et sd sont nécessaires pour finir de normaliser les résidus pour le mois de janvier après le retrait de la tendance saisonnière.

4.4.2 Simulation conditionnelle

La distribution conditionnelle dans un cadre gaussien est rappelée à la sous section 2.1.2, et est mise en application au chapitre 3 pour la simulation de variables météorologiques (rayonnement solaire, température moyenne, humidité) au Rheu et à Bourran. Les résidus en ces deux sites, modélisés par un champ gaussien, sont simulés conditionnellement aux autres résidus "observés" (on observe les variables qui sont transformées en résidus). Dans ce chapitre, la principale différence provient de l'ajout des précipitations au modèle. En effet, pour les autres variables, nous possédons toujours les résidus "observés" correspondant aux valeurs du champ gaussien. Ce n'est pas toujours le cas pour les précipitations puisqu'un nombre important de valeurs gaussiennes sont censurées (celles correspondant aux jours secs). Or nous avons besoin de ces valeurs pour procéder à la simulation conditionnelle. Les deux étapes, que sont la simulation conditionnelle et la transformation des résidus en variables, sont donc précédées d'une étape 0 qui consiste à fournir des valeurs du champ gaussien latent en cas de censure. Pour être cohérentes, ces valeurs doivent respecter la structure spatio-temporelle du champ gaussien et être in-

férieures au seuil $y_{0,\alpha}$ correspondant. L'application de la simulation conditionnelle, présentée dans la section suivante, a été effectuée en conservant les données d'Avignon et en considérant les 17 autres sites comme conditionnants.

Prenons l'exemple d'un jour t où il n'a plu sur aucun des 17 sites. Nous voulons obtenir les valeurs gaussiennes $y_p(\mathbf{s}_\alpha, t)$, $\alpha = 1, \dots, 17$. La première méthode consiste à utiliser la loi conditionnelle

$$\pi(y_p(\mathbf{s}_1, t), \dots, y_p(\mathbf{s}_{17}, t) | y_1(\mathbf{s}_1, t), \dots, y_{p-1}(\mathbf{s}_1, t), y_1(\mathbf{s}_2, t), \dots, y_{p-1}(\mathbf{s}_{17}, t)), \quad (4.19)$$

en vérifiant la condition relative au seuil pour chacune des 17 valeurs obtenues. Tant que cette condition n'est pas satisfaite, de nouvelles valeurs sont générées suivant la loi conditionnelle 4.19. Cependant, le temps d'obtention de valeurs cohérentes risque d'être long pour certaines configurations. Par exemple, nous avons eu le cas d'un jour où il pleuvait partout, et avec forte intensité, sauf en 2 sites (Auzeville-Tolosane et Lavaur) ; dans cette situation, cet algorithme a été particulièrement long pour fournir les deux valeurs gaussiennes censurées.

Lorsque le nombre de stations est élevé, nous recommandons l'utilisation d'un échantillonneur de Gibbs, comme dans [Allcroft and Glasbey \[2003\]](#) et [Leblois \[2012\]](#). L'échantillonneur de Gibbs permet d'obtenir successivement les valeurs $y_p(\mathbf{s}_\alpha, t)$ à l'aide de distributions conditionnelles univariées. La première distribution conditionnelle s'écrit :

$$\pi(y_p(\mathbf{s}_1, t) | y_1(\mathbf{s}_1, t), \dots, y_{p-1}(\mathbf{s}_1, t), y_1(\mathbf{s}_2, t), \dots, y_{p-1}(\mathbf{s}_{17}, t)). \quad (4.20)$$

La valeur $y_p(\mathbf{s}_1, t)$ est retenue dès qu'elle satisfait à la condition $y_p(\mathbf{s}_1, t) \leq y_{0,1}$. Puis la seconde distribution intègre dans son conditionnement la nouvelle valeur de $y_p(\mathbf{s}_1, t)$ ce qui se traduit par :

$$\pi(y_p(\mathbf{s}_2, t) | y_1(\mathbf{s}_1, t), \dots, y_{p-1}(\mathbf{s}_1, t), y_p(\mathbf{s}_1, t), y_1(\mathbf{s}_2, t), \dots, y_{p-1}(\mathbf{s}_{17}, t)), \quad (4.21)$$

et ainsi de suite pour les sites suivants. Ces 17 premières valeurs générées, nommées $(y_p^0(\mathbf{s}_1, t), \dots, y_p^0(\mathbf{s}_{17}, t))$, servent de valeurs initiales pour la suite de l'algorithme. Il faut ensuite itérer un grand nombre de fois (plusieurs centaines au moins), la loi conditionnelle univariée

$$\pi(y_p^{m+1}(\mathbf{s}_\alpha, t) | y_1(\mathbf{s}_1, t), \dots, y_p^{m+1}(\mathbf{s}_{\alpha-1}, t), y_1(\mathbf{s}_\alpha, t), \dots, y_{p-1}(\mathbf{s}_\alpha, t), y_1(\mathbf{s}_{\alpha+1}, t), \dots, y_p^m(\mathbf{s}_{17}, t)), \quad (4.22)$$

pour $\alpha = 1, \dots, 17$, afin d'obtenir la convergence de l'algorithme vers la loi stationnaire. Comme précédemment, la valeur $y_p^{m+1}(\mathbf{s}_\alpha, t)$ est retenue lorsqu'elle satisfait la condition $y_p^{m+1}(\mathbf{s}_\alpha, t) \leq y_{0,\alpha}$, et nous pouvons ainsi actualiser le conditionnement.

Par soucis de simplicité, nous avons écrit uniquement le cas spatial, c'est à dire sans différence de temps dans le conditionnement, mais le cas spatio-temporel est similaire.

4.5 Validation du générateur stochastique de temps

4.5.1 Simulation non conditionnelle

L'intérêt principal d'un SWG est de générer des séquences aussi vraisemblables que celles observées. Pour valider la qualité des séquences simulées, nous nous sommes appuyés sur un certain nombre de statistiques (densité, durée d'événements, corrélations par paires de sites, corrélation par paires de variables,...). Il en existe bien d'autres et dans un contexte spatio-temporel multivarié, il serait illusoire de vouloir être exhaustif.

Les données historiques sont les 11 mois de janvier 2003 à 2013. Nos simulations sont construites de la même façon puisque 11 périodes de 31 jours sont générées. Cette opération est répétée 100 fois. Les comparaisons de statistiques se font entre les observations et les données simulées (i.e. le processus $\mathbf{Z}$), et non les résidus (i.e. le processus $\mathbf{Y}$).

Les figures 4.8, 4.9, 4.10 et 4.11 montrent respectivement les densités observées (en noir) et simulées (en gris) des variables T_n, T_x, V et R pour chacun des 18 sites. Les températures minimales et maximales sont plutôt bien reproduites pour l'ensemble des sites, notamment pour la position du mode, sauf pour les températures maximales à Auzerville-Tolosane et Lavar. Ce décalage du mode pour ces deux stations - qui sont hors zone méditerranéenne - provient probablement des paramètres saisonniers de la moyenne pour les jours pluvieux (plus nombreux à l'ouest). Les deux variables suivantes, V et R , présentent une forte asymétrie. Le rayonnement présente en plus une bimodalité très marquée (cf section 4.2). La construction du modèle permet de gérer avec plus ou moins de réussite ces difficultés. La transformation log appliquée aux vitesses de vent (ou exp aux résidus correspondants) permet de restituer pour la plupart des sites la forme des densités avec cette asymétrie caractéristique. Certains sites comme Montélimar présentent des densités simulées moins en adéquation avec celle observée. Le choix d'une transforme unique, correspondant à $\lambda = 0$ pour l'équation de boxcox 4.4, explique ces écarts. Une valeur de λ différente aurait été souhaitable au prix d'une complexification, car nous utilisons la transformation log pour tous les mois et toutes les saisons quelque soit le site. Nous observons également une bonne adéquation des densités pour le mois de juillet. L'utilisation de paramètres saisonniers - moyenne et écart-type - différents selon l'occurrence de précipitation apporte une certaine flexibilité et autorise le modèle à reproduire une partie de l'asymétrie et de la bimodalité de la densité du rayonnement solaire. Cependant cela n'est pas entièrement satisfaisant, notamment pour les sites où la bimodalité est très marquée comme Avignon. La partition selon les précipitations n'est sans doute pas suffisante, et de plus nombreux états de temps, comme dans [Flecher et al., 2010], permettraient sans doute une meilleur adéquation.

La capacité du modèle à reproduire correctement les intensités de précipitation a déjà

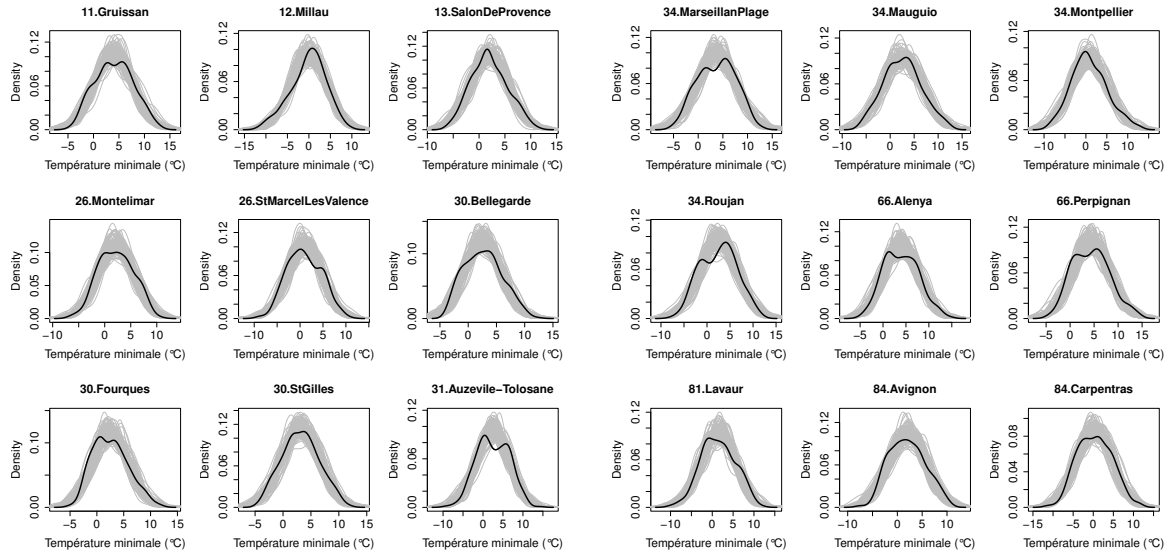


FIGURE 4.8 – Représentation de la densité par la méthode du noyau. En noir : les températures minimales observées pour les 11 mois de janvier (2003-2013) aux 18 sites. En gris : 100 simulations.

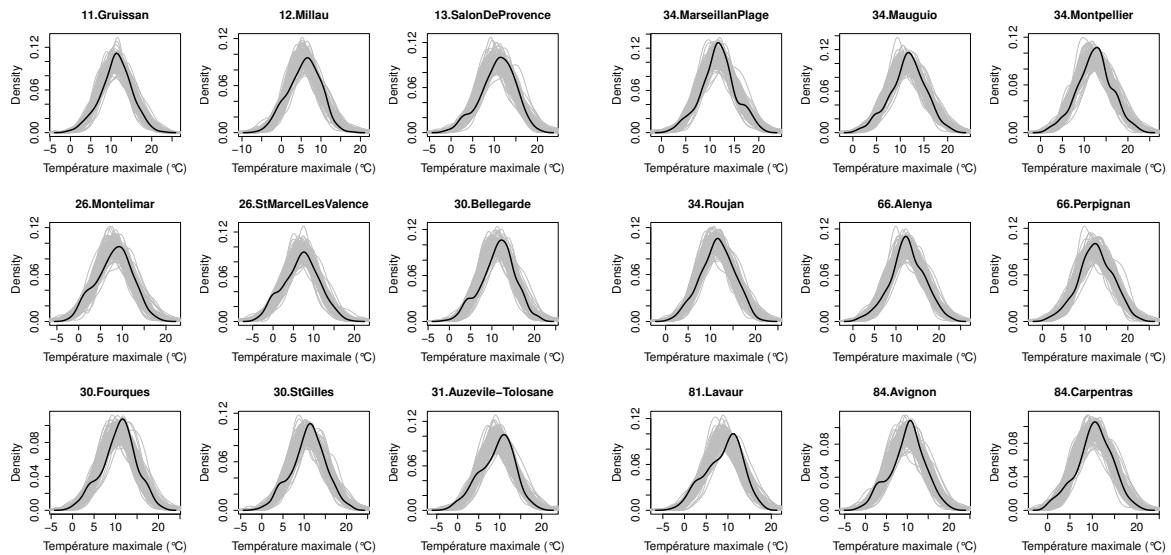


FIGURE 4.9 – Représentation de la densité par la méthode du noyau. En noir : les températures maximales observées pour les 11 mois de janvier (2003-2013) aux 18 sites. En gris : 100 simulations.

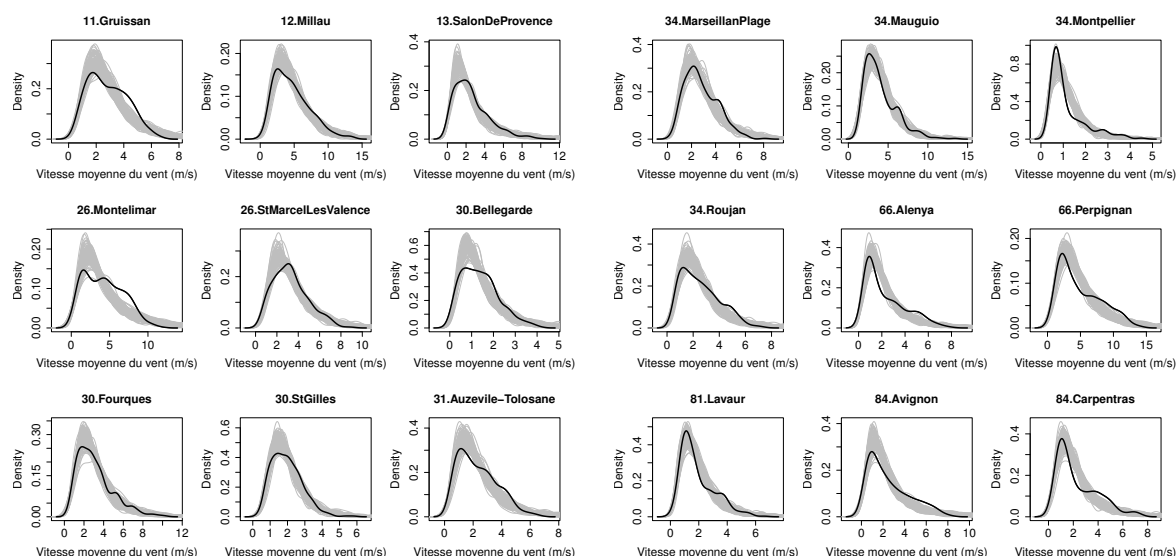


FIGURE 4.10 – Représentation de la densité par la méthode du noyau. En noir : les vitesses moyennes de vent observées pour les 11 mois de janvier (2003-2013) aux 18 sites. En gris : 100 simulations.

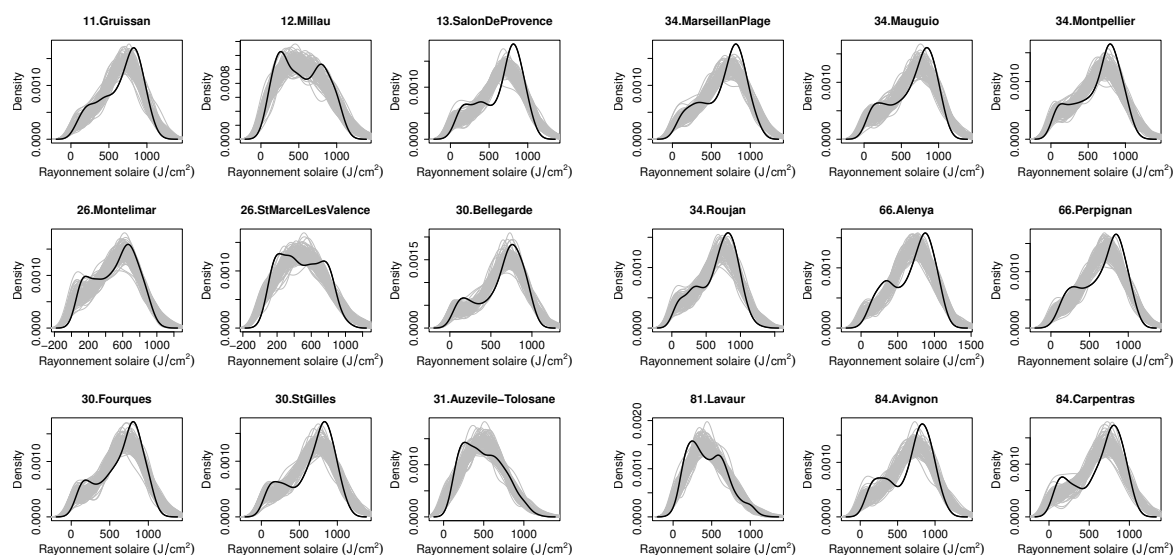


FIGURE 4.11 – Représentation de la densité par la méthode du noyau. En noir : le rayonnement solaire observé pour les 11 mois de janvier (2003-2013) aux 18 sites. En gris : 100 simulations.

été discutée dans la sous-section 4.3.2 et dans Allard and Bourotte [2014]. Nous allons maintenant nous intéresser à la distribution des durées d'événements secs et pluvieux puisque les précipitations sont caractérisées par leur intensité et leur occurrence. La figure 4.12 offre une visualisation du nombre d'événements pluvieux en fonction de leur durée (en jours). Les observations sont en rouge. Il y a assez peu d'événements de plus de 5 jours sur l'ensemble des sites. Les boxplots calculés à partir des simulations démontrent la capacité du champ gaussien latent à simuler des durées d'événements pluvieux réalistes. Seule la station d'Auzeville-Tolosane pose soucis avec une sous estimation du nombre d'événements pluvieux d'un jour, ce qui n'est pas le cas avec la station de Lavaur. De plus, un événement extrême (en durée) de 20 jours consécutifs pluvieux a été enregistré, ce que le modèle n'est pas capable de reproduire (au mieux 13 jours consécutifs). La région méditerranéenne est notamment caractérisée par son climat sec. Il est possible de calculer les distributions des événements secs à l'instar des événements pluvieux, et c'est ce qui est présenté par la figure 4.13. Nous pouvons remarquer que le modèle reproduit bien cette statistique. En particulier, il est capable de simuler des événements secs particulièrement longs de 15 à 20 jours pour la plupart des sites, où ce type d'événements se produit. En revanche, au delà de 20 jours, le modèle peine à générer des événements secs de durée extrême (par exemple, un mois entier sans pluie à Bellegarde). Il serait intéressant de travailler sur une saison pour voir si ces durées très longues peuvent être reproduites.

Deux autres statistiques sur les précipitations ont été étudiées. La première, appelée cumul mensuel agrégé, vérifie que le comportement de l'ensemble des stations à l'échelle de la région soit cohérent. Pour se faire, nous calculons la somme régionale des cumuls de précipitation pour chacun des 11 mois de janvier dont nous disposons. La moyenne et l'écart-type des cumuls mensuels agrégés sont ensuite obtenus pour ces 11 répétitions. La même procédure est effectuée sur les 100 simulations. Les résultats sont résumés dans le tableau 4.5. La moyenne des cumuls mensuels agrégés est très bien reproduite par le modèle. En revanche, le modèle produit plus de variabilité qu'observée sur les données.

La seconde statistique donne une indication sur le comportement des différents sites lors d'événements de pluie extrême. Ainsi, nous nous sommes intéressés au nombre de stations qui dépassent en même temps un certain seuil de précipitations. Ce seuil peut être considéré comme la limite entre les pluies classiques et les extrêmes. Il a été déterminé en pratiquant un *mean residual life plot* (fonction *mrl.plot* du package *ismev*) pour chacune des stations et la valeur retenue est 8 mm. Les résultats, présentés par la figure 4.14, montrent une surestimation du nombre d'événements extrêmes isolés, c'est à dire ne se produisant qu'en un seul site. Pour les événements de pluie extrême se produisant sur plusieurs sites, les simulations témoignent d'un comportement très fidèle aux observations.

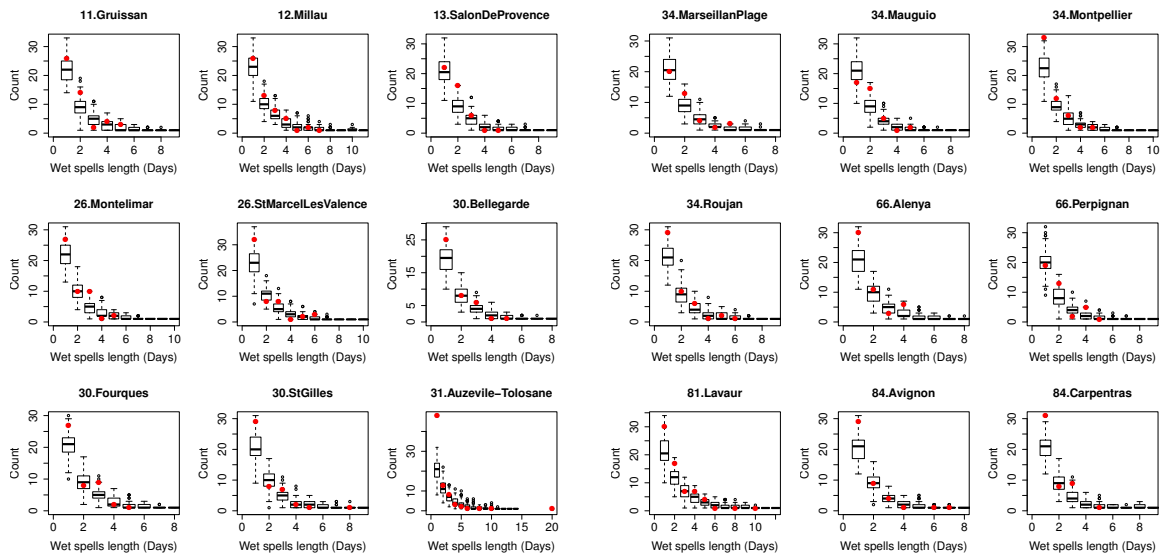


FIGURE 4.12 – Nombre d'événements pluvieux en fonction de leur durée (en jours). En rouge : le compte observé pour les 11 mois de janvier (2003-2013) aux 18 sites. Les boxplots sont calculés à partir des 100 simulations.

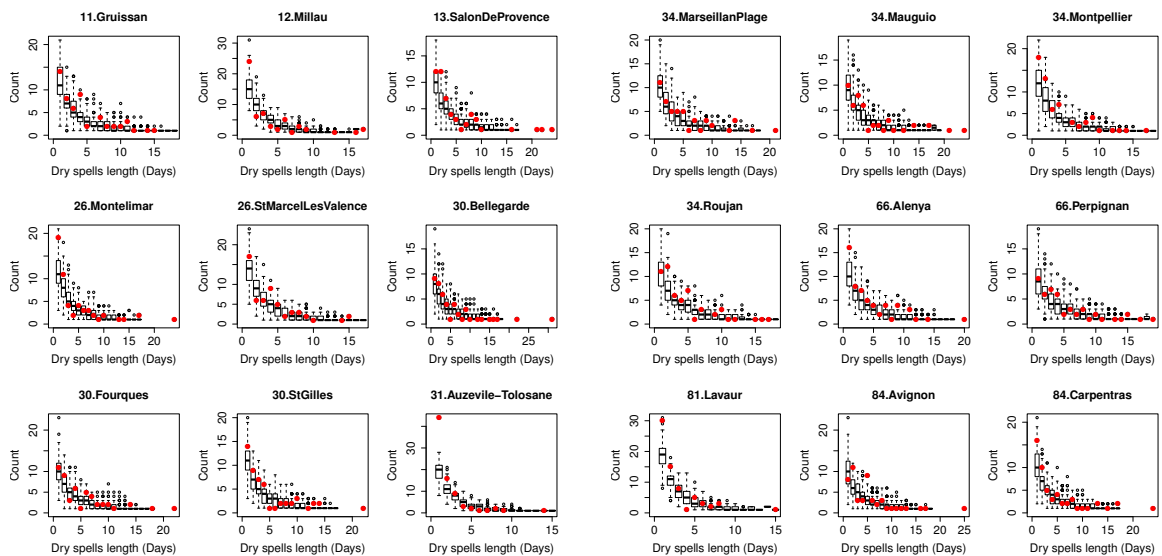


FIGURE 4.13 – Nombre d'événements secs en fonction de leur durée (en jours). En rouge : le compte observé pour les 11 mois de janvier (2003-2013) aux 18 sites. Les boxplots sont calculés à partir des 100 simulations.

	True	Min	Q ₁	Median	Mean	Q ₃	Max
moyenne	881	281	732	943	948	1085	1988
écart-type	593	190	625	756	913	1083	3484

TABLE 4.5 – Statistiques sur les cumuls mensuels agrégés des précipitations de la zone d'étude pour les mois de janvier simulés. La valeur *true* correspond aux cumuls observés sur la période 2003-2013.

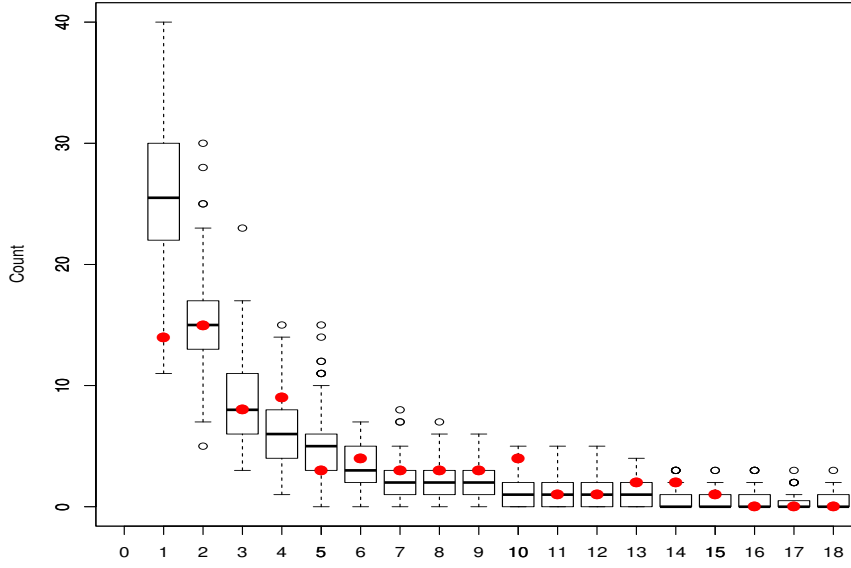


FIGURE 4.14 – Nombre de stations dépassant simultanément le seuil de précipitation 8 mm. Les points rouges correspondent aux données observées et les boxplots aux 100 simulations.

La figure 4.15 est composée de 18 graphiques que l'on appelle des graphiques en étoile. Pour chaque graphique en étoile, il y a dix sommets correspondant aux dix paires de variables ($\frac{p(p-1)}{2}$ avec $p = 5$). Les résultats sont satisfaisants puisque dans la plupart des cas, le tracé noir (observations) est dans l'enveloppe grise (simulations). Nous pouvons remarquer que d'un site à l'autre, le tracé noir est relativement variable. Ce tracé symbolise les éléments de la matrice de corrélation, hors diagonale. L'enveloppe grise est moins variable d'un site à l'autre car le processus gaussien $\mathbf{Y}$, qui est ensuite transformé en processus $\mathbf{Z}$ (tracé gris), possède la même matrice de corrélation $\left[\sigma_{ij}^{(Y)}\right]_{i,j=1}^p$ pour tous les sites (cf tableau 4.4 pour les estimations de ces paramètres).

4.5.2 Simulation conditionnelle

Un des intérêts majeurs de notre générateur stochastique de temps est de permettre la simulation conditionnelle. Dans cet exemple, nous avons généré les 5 variables météorologiques pour la station d'Avignon sur tout le mois de janvier 2014 conditionnellement aux 17 autres stations. Les algorithmes de simulation, présentés en 4.4.2, ont un fonctionnement séquentiel, et simulent les jours les uns après les autres. Les résultats présentés ici ont été obtenus à l'aide de la première méthode (distribution conditionnelle 4.19) pour générer les valeurs censurées. Pour le jour t , seuls les deux jours précédents $t - 2$ et $t - 1$ sont pris en compte sur l'ensemble des stations ainsi que les données des 17 autres stations au jour t . Nous utilisons les paramètres saisonniers et les paramètres de la fonction d'anamorphose ψ estimés sur les mois de janvier 2003 à 2013. Cette situation pourrait

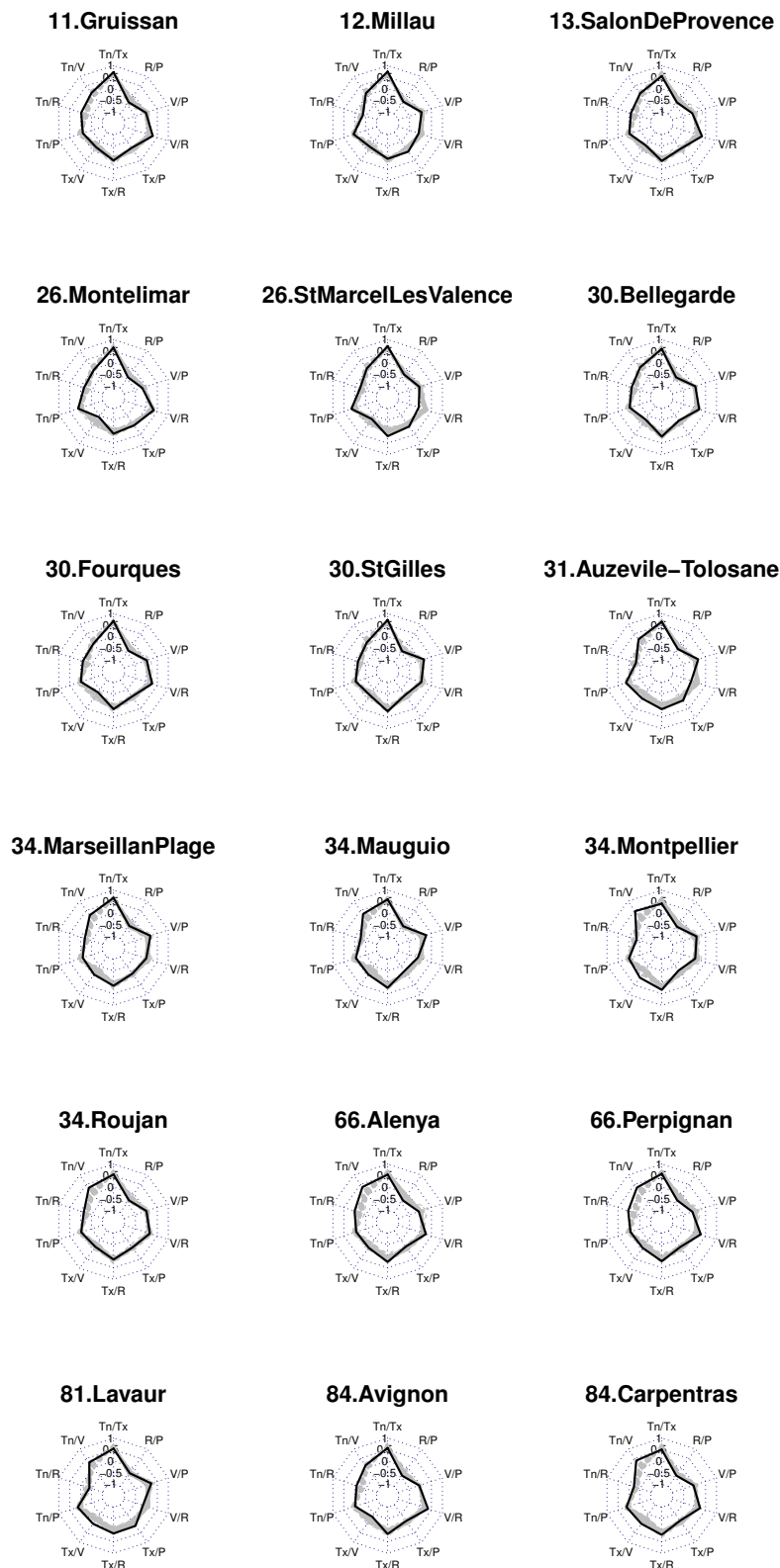


FIGURE 4.15 – Corrélations entre chaque paire de variables pour les 18 stations. En noir : sur les données observées pour les 11 mois de janvier (2003-2013). En gris : sur les 100 simulations.

correspondre à une panne sur la station d'Avignon pour le mois de janvier 2014.

Nous avons réalisé 100 simulations conditionnelles pour la station d'Avignon et les résultats sont représentés par les différents graphiques des figures 4.16 et 4.17. Ces résultats plutôt satisfaisants amènent un certain nombre de commentaires. Le premier concerne la ligne pointillée rouge. Contrairement à la figure 3.8 du chapitre 3, cette ligne rouge pointillée ne correspond pas à l'espérance conditionnelle mais à la moyenne des 100 simulations. En effet, d'une simulation à l'autre, l'espérance conditionnelle peut varier puisque les valeurs conditionnantes changent pour les valeurs censurées (première étape de l'algorithme).

L'allure générale des températures minimales est bien reproduite mais la plupart des valeurs observées sont au dessus de notre valeur moyenne (ligne rouge). Cet phénomène est particulièrement notable pour les températures observées supérieures à 6°C. Si l'on se réfère à la densité des températures minimales 4.8 à Avignon, nous pouvons remarquer que la génération de valeurs entre 6°C et 10°C est légèrement sous estimée, ce qui expliquerait ces observations. De plus, la douceur du mois de janvier 2014 est remarquable puisque c'est le plus chaud en France depuis 1900 (<http://www.meteofrance.fr/climat-passe-et-futur/bilans-climatiques/bilan-2014/bilan-climatique-de-l-hiver-2013-2014>). La modélisation saisonnière pourrait intégrer une tendance linéaire. Ceci n'a pas été testé ici puisque les problématiques de réchauffement climatique sont en dehors du champ d'étude de cette thèse. Les températures maximales simulées sont plutôt fidèles à celles observées.

Les variations des vitesses moyennes simulées du vent sont très cohérentes avec les observations. Cependant, un biais conduit en moyenne à des vitesses simulées légèrement trop hautes. Il est possible que la transformation log ($\lambda = 0$), choisie pour toute la région, ne soit pas la plus adaptée pour Avignon (i.e. $\lambda > 0$).

Malgré la bimodalité (figure 4.2) du rayonnement solaire, la valeur moyenne est proche des valeurs observées, faibles comme hautes. Cependant les valeurs les plus élevées, des jours 20 à 27, sont sous estimées ce qui est cohérent avec les observations faites sur le graphique 4.11 des densités. Le mode le plus haut est sous estimé pour Avignon. De plus sur cette période (jours 20 à 27), des valeurs faibles appartiennent à l'enveloppe de confiance (90%). Ceci doit être analysé conjointement avec la figure 4.17 puisqu'une proportion non négligeable (entre 10 et 50% selon les jours) de jours pluvieux ont été simulés à tort. Les jours pluvieux possèdent une moyenne saisonnière plus faible pour le rayonnement solaire ce qui explique la courbe grise.

Enfin, le mois de janvier 2014 présente quelques particularités aussi pour les précipitations (<http://www.meteofrance.fr/climat-passe-et-futur/bilans-climatiques/bilan-2014/bilan-climatique-de-l-hiver-2013-2014>). Le cumul des précipitations est de 50 à 100% supérieur aux moyennes de janvier pour la région PACA, en revanche il est de 50% en déficit pour la zone de Montpellier à Perpignan. La proportion de jours secs pour Avignon en

janvier 2014 est de 58% contre 78% sur la période d'étude (tableau 4.1). Sur nos 100 simulations, la proportion de jours secs est bien restituée puisqu'elle est de 57% (avec [46,68] comme intervalle de confiance à 95%). Jusqu'au jour 20 (figure 4.17), les simulations sont très satisfaisantes. La zone d'étude se comporte de manière homogène sur cette période.

A partir du 21 ième jour, des comportements différents apparaissent entre la zone à l'ouest et à l'est de Nîmes. Par exemple pour le 23, il ne pleut pas sur Avignon comme sur les stations proches telles que Salon de Provence, Carpentras ou Fourques. Pourtant, le modèle produit pour 20% des simulations un jour pluvieux sur Avignon. Des précipitations sont enregistrées sur les stations de Perpignan à Montpellier ce jour là, ce qui augmente par le jeu du conditionnement la valeur du potentiel de pluie (i.e. le champ gaussien latent) à Avignon. Le pic au jour 25 est provoqué par une faible pluie enregistrée à Carpentras, la station la plus proche d'Avignon. Enfin, deux jours de fortes précipitations sont enregistrées à Avignon à la fin du mois de janvier 2014. Il est intéressant de noter que la surface grisée englobe ces forts cumuls, même si la moyenne (ligne rouge) est de seulement 20 mm. Cela s'explique encore par une différence de comportement entre les stations proches d'Avignon (forts cumuls en Camargue et PACA) et celles à l'ouest (faibles cumuls).

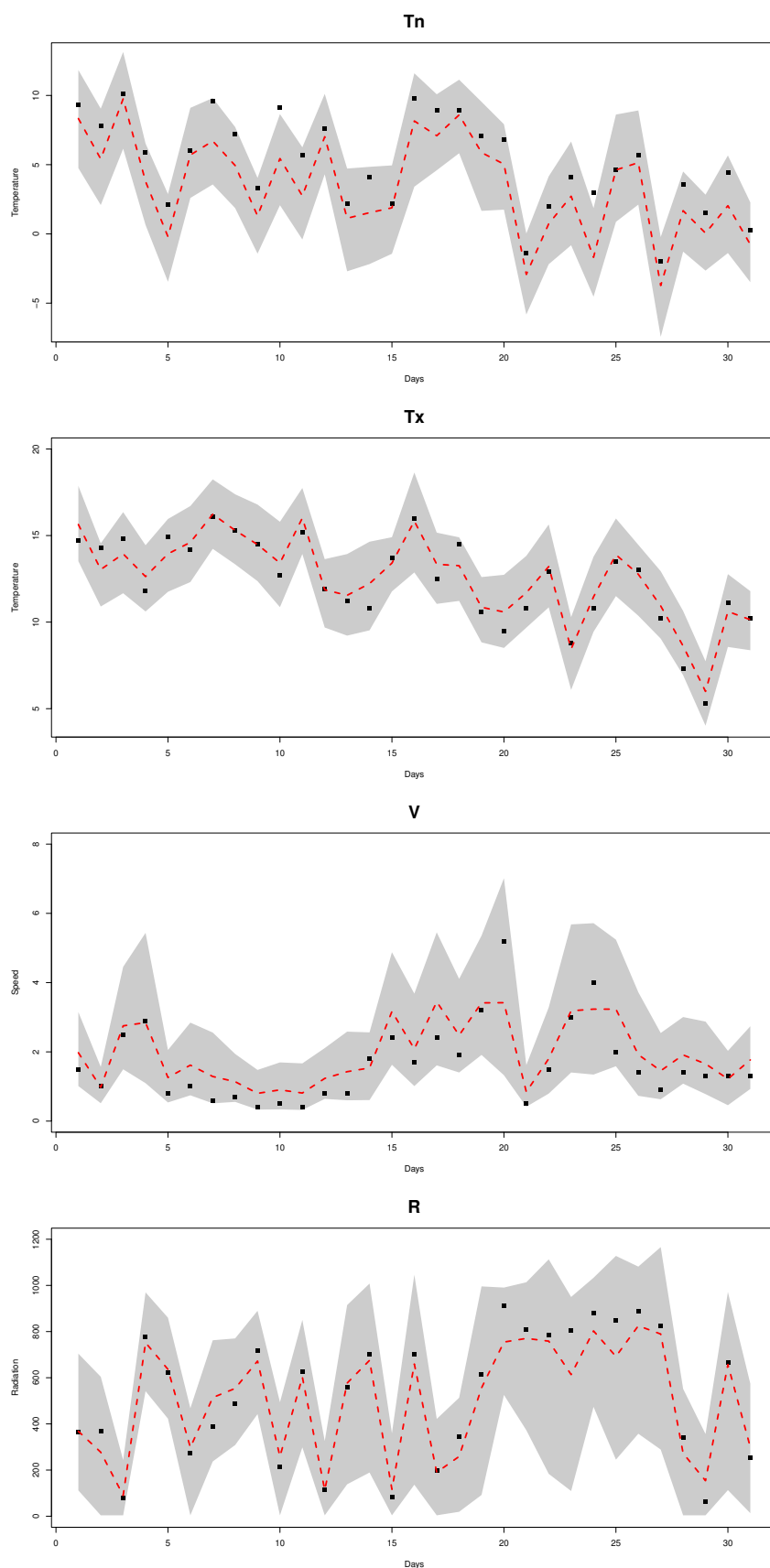


FIGURE 4.16 – Simulation conditionnelle des températures minimales et maximales, de la vitesse du vent et du rayonnement solaire à Avignon pour janvier 2014. Les carrés noirs correspondent aux données réelles. La ligne rouge pointillée symbolise la valeur moyenne des 100 simulations, et l’aire grisée 90% des valeurs simulées.

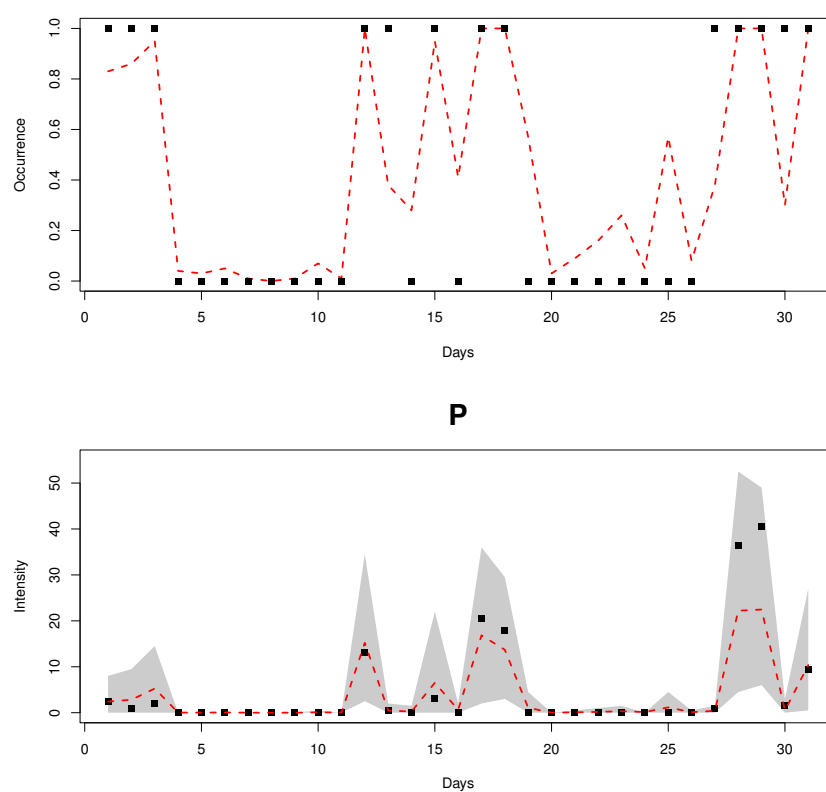


FIGURE 4.17 – Simulation conditionnelle de l'occurrence (en haut) et de l'intensité (en bas) des précipitations à Avignon pour janvier 2014. Les carrés noirs représentent les données réelles. Pour les occurrences, la valeur 1 correspond à un jour pluvieux et 0 à un jour sec. La ligne rouge pointillée symbolise la valeur moyenne des 100 simulations, et l'aire grisée 90% des valeurs simulées.

4.6 Conclusion et perspectives

Le générateur stochastique de temps développé dans ce chapitre offre une alternative intéressante aux SWGs multisites basés sur le modèle de Richardson. Notre SWG s'articule autour d'un unique champ gaussien multivarié, ce qui le rend très facile à mettre en œuvre. Les modèles avancés de covariance et covariance croisée, comme la covariance croisée entièrement non séparable utilisée ici, apportent une grande flexibilité au modélisateur tout en lui garantissant que la matrice de covariance soit semi-définie positive, quelque soit le nombre de jours, de sites ou de variables. Les approches construites autour du concept de vraisemblance composite poursuivent leur développement en parallèle, et fournissent une estimation rapide et efficace des paramètres de la fonction de covariance, même pour de grands jeux de données.

Nous avons testé notre modèle sur une zone d'étude à climat méditerranéen au travers de simulations conditionnelles et non conditionnelles des températures minimales et maximales, de la vitesse du vent, du rayonnement solaire et des précipitations. Les résultats obtenus sont cohérents et encourageants, même si il reste certaines statistiques qui ne sont pas encore reproduites correctement par notre SWG. La fonction d'anamorphose choisie pour les précipitations semble particulièrement pertinente. Plusieurs pistes d'amélioration sont en réflexion.

La première mène naturellement au concept d'état de temps qui est de plus en plus utilisé pour les générateurs stochastiques de temps [Boé et al., 2006] [Boé and Terray, 2008]. A l'intérieur d'un même état de temps, les différents jours possèdent des propriétés statistiques similaires. Les états de temps peuvent être déterminés à l'aide d'une classification [Flecher et al., 2010] ou être considérés comme cachés. Ils sont alors modélisés par une variable latente [Zucchini and Guttorp, 1991] [Ailliot and Monbet, 2012]. Les premiers états de temps se sont basés sur la dichotomie jours secs et jours pluvieux, mais les travaux récents amènent souvent à un nombre un peu plus élevé d'états de temps (entre 3 et 6 suivant les saisons dans Flecher et al. [2010]). Les états de temps permettent de gérer les variables à densités asymétriques - telles que le vent, le rayonnement solaire - sans avoir à effectuer de transformation pour les rendre gaussienne [Ailliot and Monbet, 2012]. Dans notre modèle, nous avons uniquement différencié les paramètres saisonniers en fonction des précipitations. Même si cela apporte une amélioration de l'adéquation des densités, ce n'est pas suffisant pour le rayonnement solaire. De plus, cette opération s'opère sur la structure d'ordre 1 du champ, tandis que la structure d'ordre 2 reste commune aux jours pluvieux et secs.

L'introduction d'états de temps plus complexes dans le modèle par l'intermédiaire d'une variable cachée semble compliquée. En effet, il faudrait revoir entièrement le modèle en le transformant en modèle auto-régressif. Ensuite l'estimation, dans ce type de modèle, se fait souvent par vraisemblance maximisée par un algorithme EM (espérance-

maximisation). Il semble plus simple de s'orienter vers une méthode de classification pour les mélanges gaussiens comme Mclust [Fraleigh and Raftery, 2006], utilisé par Fletcher et al. [2010] pour le générateur WACS-Gen. Ainsi à l'intérieur d'un état de temps, notre modèle basé sur un champ gaussien multivarié serait opérationnel, tout comme la méthode d'estimation. D'un état à l'autre, seuls les paramètres changeraient (notamment ceux du modèle de covariance) et les transitions entre les états de temps seraient modélisées par une chaîne de Markov. Cependant, à la différence de WACS-Gen, notre générateur est multisite ce qui nécessite de prendre en considération la dimension spatiale pour l'établissement des états de temps. Cette question méthodologique reste ouverte. Une option serait de supposer les états de temps comme régionaux, ou régime de temps, pour éviter d'avoir à introduire une structure spatiale. La classification de type Mclust s'effectuerait en considérant les x variables aux y sites comme xy variables. Cette option est valable pour un nombre restreint de sites et de variables.

La seconde piste d'amélioration conduit à développer des modèles de covariance intégrant mieux la physique des phénomènes observés. Comme nous l'avons expliqué au chapitre 2, le modèle de covariance croisée entièrement non séparable développé dans cette thèse est une brique nécessaire pour construire des modèles plus flexibles capables de prendre en compte la complexité des données observées.

Une première étape serait d'intégrer une anisotropie géométrique pour la partie spatiale du modèle. Pour $d = 2$ et u fixé, la fonction de covariance croisée a des contours de forme circulaire. L'anisotropie géométrique permet aux contours d'avoir une forme elliptique dans une direction privilégiée. Cela se fait en appliquant aux coordonnées une transformation linéaire. La distance euclidienne est alors remplacée par :

$$\|A_{\theta, \epsilon} \mathbf{h}\| = \sqrt{\epsilon (h_1 \cos \theta + h_2 \sin \theta)^2 + \frac{1}{\epsilon} (-h_1 \sin \theta + h_2 \cos \theta)^2}, \quad (4.23)$$

où θ correspond à l'angle de rotation et ϵ le coefficient de dilatation. Le cas isotrope est obtenu pour $\epsilon = 1$. Baxevani and Lennatsson [2015] ont utilisé cette distance pour leur générateur de pluie basé sur un champ gaussien latent. Une application de l'anisotropie géométrique avec une covariance de Matérn est présentée dans Haskard [2007].

Cependant, il nous semble encore plus intéressant de prendre en compte la notion de mouvement, surtout pour une région si propice aux vents violents (phénomène d'advection). Le but est de s'affranchir de l'hypothèse restrictive de symétrie, rarement réaliste pour des variables météorologiques. Pour modéliser le déplacement de cellules orageuses, Cox and Isham [1988] ont utilisé la covariance lagrangienne spatio-temporelle

$$C(\mathbf{h}, u) = \mathbb{E}[C_S(\mathbf{h} - \mathbf{V}u)], \quad (4.24)$$

avec C_S une covariance spatiale valide et $\mathbf{V}$ un vecteur aléatoire (l'espérance se réfère à ce vecteur). Ce modèle généralise le modèle figé (*frozen field model*) correspondant à $\mathbf{V} = \mathbf{v}$

constant. L'échelle spatiale doit être en cohérence avec l'échelle temporelle. La construction de covariance lagrangienne à partir de covariance spatio-temporelle valide C^0 se fait à l'aide d'un mouvement Φ (ou difféomorphisme) appartenant à une famille de mouvements présentant un certain nombre de propriétés [Ailliot et al., 2011]. Ce mouvement est lié à un champ de vélocité (ou d'advection [Leblois and Creutin, 2013]). La covariance d'un champ gaussien Y est alors de la forme :

$$C_Y((\mathbf{s}, t), (\mathbf{s}', t')) = C^0(\Phi^{-1}(\mathbf{s}', t_0, t') - \Phi^{-1}(\mathbf{s}, t_0, t), t' - t), \quad (4.25)$$

pour un point arbitraire $t_0 \in \mathbb{R}$ et $(\mathbf{s}, t), (\mathbf{s}', t')$ dans $\mathbb{R}^{d+1}$. Les travaux de Carrizo-Vergara [2015] montrent que l'approche par maximum de vraisemblance composite peut être une approche intéressante pour estimer les paramètres de C^0 (hors paramètres de vélocité), notamment pour les covariances spatio-temporelles non séparables de Gneiting. De plus, Carrizo-Vergara [2015] prouve que l'on peut étendre cette construction au cas multivarié du moment que l'on possède une famille de fonctions de covariances croisées $[C_{ij}^0]_{i,j=1}^P$ valide.

Enfin, une alternative simple, proposée par Gneiting et al. [2007], est de construire une covariance spatio-temporelle asymétrique C^{ASM} comme la somme d'une covariance non séparable de type Gneiting et d'une covariance lagrangienne

$$C^{\text{LGR}}(\mathbf{h}, u) = \left(1 - \frac{1}{2\nu}|h_1 - \nu u|\right)_+, \quad (4.26)$$

version simplifiée de 4.24, avec ν la vitesse de déplacement selon la première composante et $p_+ = \max(p, 0)$. Les données [Haslett and Raftery, 1989] de vitesse de vent utilisées dans cette étude montrent une forte dominance de la composante Est-Ouest (h_1). Les résultats obtenus à l'aide du modèle asymétrique sont meilleurs que ceux obtenus par le modèle non séparable seul, qui obtient lui-même de meilleures performances que le modèle séparable. La généralisation au cas multivarié C_{ij}^{ASM} pourrait se faire dans un premier temps en choisissant une version séparable de la covariance croisée lagrangienne telle que :

$$C_{ij}^{\text{ASM}}(\mathbf{h}, u) = C_{ij}^{\text{NS}}(\mathbf{h}, u) + C_{ij}^{\text{LGR}}(\mathbf{h}, u), \quad (4.27)$$

$$C_{ij}^{\text{LGR}}(\mathbf{h}, u) = a_{ij}C^{\text{LGR}}(\mathbf{h}, u), \quad (4.28)$$

avec C_{ij}^{NS} le modèle de covariance croisée développée dans cette thèse, et $A = [a_{ij}]$ une matrice définie positive. Il est fort probable que les déplacements de masses d'air, de nuages se fassent de manière différente (en vitesse et en direction) pendant le mois ou la saison étudié(e). La combinaison de régimes de temps associés à une fonction de covariance croisée, dont les paramètres de la partie non séparable et de la partie lagrangienne varieraient en fonction du régime de temps, permettraient une meilleure prise en compte des dynamiques des variables météorologiques, et des simulations plus réalistes par notre générateur stochastique de temps.

4.7 Références

- Ailliot, P., Baxevasani, A., Cuzol, A., Monbet, V., and Raillard, N. (2011). Space-time models for moving fields with an application to significant wave height fields. *Environmetrics*, 22(3) :354–369. [106](#)
- Ailliot, P. and Monbet, V. (2012). Markov-switching autoregressive models for wind time series. *Environmental Modelling & Software*, 30 :92–101. [104](#)
- Allard, D. and Bourotte, M. (2014). Disaggregating daily precipitations into hourly values with a transformed censored latent gaussian process. *Stochastic Environmental Research and Risk Assessment*, 29(2) :453–462. [73](#), [80](#), [96](#)
- Allcroft, D. J. and Glasbey, C. A. (2003). A latent gaussian markov random-field model for spatiotemporal rainfall disaggregation. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 52(4) :487–498. [73](#), [85](#), [92](#)
- Barancourt, C., Creutin, J., and Rivoirard, J. (1992). A method for delineating and estimating rainfall fields. *Water Resources Research*, 28(4) :1133–1144. [73](#)
- Baxevasani, A. and Lennatsson, J. (2015). A spatiotemporal precipitation generator based on a censored latent gaussian field. *Water Resources Research*. [73](#), [81](#), [105](#)
- Bessac, J., Ailliot, P., and Monbet, V. (2015). Gaussian linear state-space model for wind fields in the north-east atlantic. *Environmetrics*, 26(1) :29–38. [77](#)
- Bevilacqua, M. and Gaetan, C. (2014). Comparing composite likelihood methods based on pairs for spatial gaussian random fields. *Statistics and Computing*, pages 1–16. [86](#)
- Boé, J. and Terray, L. (2008). A weather-type approach to analyzing winter precipitation in france : twentieth-century trends and the role of anthropogenic forcing. *Journal of Climate*, 21(13) :3118–3133. [104](#)
- Boé, J., Terray, L., Habets, F., and Martin, E. (2006). A simple statistical-dynamical downscaling scheme based on weather types and conditional resampling. *Journal of Geophysical Research : Atmospheres*, 111(D23). [104](#)
- Bourotte, M., Allard, D., and Porcu, E. (2016). A flexible class of non-separable cross-covariance functions for multivariate space-time data. *Spatial Statistics*, <http://dx.doi.org/10.1016/j.spasta.2016.02.004>. [72](#)
- Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252. [77](#)

- Brown, B. G., Katz, R. W., and Murphy, A. H. (1984). Time series models to simulate and forecast wind speed and wind power. *Journal of climate and applied meteorology*, 23(8) :1184–1195. [77](#)
- Carrizo-Vergara, R. A. (2015). Modèles de champs aléatoires spatio-temporels non séparables et non symétriques : étude théorique et estimation des paramètres de mouvement. Master's thesis. [106](#)
- Chilès, J. and Delfiner, P. (1999). *Geostatistics : Modeling Spatial Uncertainty*. New York : Wiley. [73](#)
- Cox, D. and Isham, V. (1988). A simple spatial-temporal model of rainfall. In *Proceedings of the Royal Society of London A : Mathematical, Physical and Engineering Sciences*, volume 415, pages 317–328. The Royal Society. [105](#)
- Flecher, C., Naveau, P., Allard, D., and Brisson, N. (2010). A stochastic daily weather generator for skewed data. *Water Resources Research*, 46(7) :W07519. [72](#), [77](#), [93](#), [104](#), [105](#)
- Fraley, C. and Raftery, A. E. (2006). Mclust version 3 : an r package for normal mixture modeling and model-based clustering. Technical report, DTIC Document. [105](#)
- Genton, M. G. and Kleiber, W. (2014). Cross-covariance functions for multivariate geostatistics. *Statistical Science*. [72](#)
- Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458) :590–600. [83](#)
- Gneiting, T., Genton, M., and Guttorp, P. (2007). Geostatistical space-time models, stationarity, separability and full symmetry. *Statistical Methods for Spatio-Temporal Systems*, pages 151–175. [106](#)
- Guillot, G. and Lebel, T. (1999). Approximation of sahelian rainfall fields with metagaussian random functions. *Stochastic Environmental Research and Risk Assessment*, 13(1-2) :113–130. [73](#)
- Haskard, K. A. (2007). An anisotropic matern spatial covariance model : Reml estimation and properties. [105](#)
- Haslett, J. and Raftery, A. E. (1989). Space-time modelling with long-memory dependence : Assessing ireland's wind power resource. *Applied Statistics*, pages 1–50. [77](#), [106](#)
- Khalili, M., Brissette, F., and Leconte, R. (2009). Stochastic multi-site generation of daily weather data. *Stochastic environmental research and risk assessment*, 23(6) :837–849. [72](#)

- Khalili, M., Leconte, R., and Brissette, F. (2007). Stochastic multisite generation of daily precipitation data using spatial autocorrelation. *Journal of hydrometeorology*, 8(3) :396–412. [72](#)
- Leblois, E. (2012). Le bassin versant, système spatialement structuré et soumis au climat. [92](#)
- Leblois, E. and Creutin, J.-D. (2013). Space-time simulation of intermittent rainfall with prescribed advection field : Adaptation of the turning band method. *Water Resources Research*, 49(6) :3375–3387. [73](#), [106](#)
- Rajagopalan, B. and Lall, U. (1999). A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water resources research*, 35(10) :3089–3101. [72](#)
- Richardson, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1) :182–190. [72](#), [77](#), [78](#)
- Richardson, C. W. and Wright, D. A. (1984). *WGEN : A model for generating daily weather variables*. US Department of Agriculture, Agricultural Research Service Washington, DC, USA. [72](#)
- Rychlik, I. (2015). Spatio-temporal model for wind speed variability. *Annales de l'I.S.U.P.*, 59 :25–56. [77](#), [83](#)
- Semenov, M. A. and Barrow, E. M. (1997). Use of a stochastic weather generator in the development of climate change scenarios. *Climatic change*, 35(4) :397–414. [72](#)
- Semenov, M. A., Brooks, R. J., Barrow, E. M., and Richardson, C. W. (1998). Comparison of the wgen and lars-wg stochastic weather generators for diverse climates. *Climate research*, 10(2) :95–107. [72](#)
- Wilks, D. (1998). Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210(1) :178–191. [72](#)
- Wilks, D. (1999). Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain. *Agricultural and Forest meteorology*, 96(1) :85–101. [72](#), [78](#)
- Zucchini, W. and Guttorp, P. (1991). A hidden markov model for space-time precipitation. *Water Resources Research*, 27(8) :1917–1923. [104](#)

Annexe A

Annexes

Article Allard & Bourotte 2014

Disaggregating daily precipitations into hourly values with a transformed censored latent Gaussian process

Denis Allard · Marc Bourotte

Published online: 2 July 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract A problem often encountered in agricultural and ecological modeling is to disaggregate daily precipitations into vectors of hourly precipitations used as input values by crop and plant models. A stochastic model for rainfall data, based on transformed censored latent Gaussian process is described. Compared to earlier similar work, our transform function provides an accurate fit for both the body and the heavy tail of the precipitation distribution. Simple empirical relationships between the parameters estimated at different time scales are established. These relationships are used for the disaggregation of daily values at stations where hourly values are not available. The method is illustrated on two stations located in the Paris basin.

Keywords Stochastic weather generator · Composite likelihood · Pairwise likelihood · Truncated Gaussian process

1 Introduction

Many crop and plant models used in agriculture and plant pathology require meteorological data at the hourly time scale as inputs (Huber et al. 1992; Caubel et al. 2012). Since hourly variables are costly to measure, record and archive, it is often the case that only daily records are available. A problem often encountered in agricultural and ecological modeling is thus to disaggregate daily values

into vectors of hourly values. Hourly temperatures can easily be synthesized from daily maximum and minimum temperatures. For precipitation, disaggregation is less straightforward. In temperate regions such as Western Europe, to account for the random occurrence of precipitations during the day, temporal disaggregation of daily rainfall is best addressed with stochastic approaches.

There is a growing literature on stochastic precipitation generators and stochastic weather generators. They are now key features of climate impact studies, particularly in hydrology and agriculture. Stochastic weather generators are statistical models that aim at simulating quickly and realistically random sequences of atmospheric variables, such as precipitation, temperature, radiation, wind speed and relative humidity. Introduction and history of stochastic weather generators, for which precipitation is one component, have been covered by Wilks and Wilby (1999), Wilks (2010) and Ailliot et al. (submitted). Historically, rainfall occurrence has been described by Markov chains (Katz 1977; Lennartsson et al. 2008; Chen et al. 2012), including non-homogeneous Markov chains (Katz and Parlange 1995; Furrer and Katz 2007; Ailliot and Monbet 2012), semi-Markov models (Racsko et al. 1991; Semenov et al. 1998) and multi-state Markov chains (Flecher et al. 2010). Another approach is to consider hidden Markov models (HMM) for occurrence (see e.g. Hughes et al. 1999; Zheng and Katz 2008). In all cases, interpretability of the states remains a key feature of these models. During wet days, a large class of distributions can be fitted to rainfall amounts. The Gamma (Richardson 1981; Flecher et al. 2010; Kleiber et al. 2013) and power transform (Ailliot et al. 2009; Zheng and Katz 2008) are among the most widely used transform functions.

Another type of approach (see e.g. Allcroft and Glasbey 2003; Thompson et al. 2007; Ailliot et al. 2009; Kleiber et al. 2013) is to define a latent Gaussian variable for which

D. Allard (✉) · M. Bourotte
UR546 Biostatistics and Spatial Processes (BioSP), INRA,
Site Agroparc, 84914 Avignon, France
e-mail: allard@avignon.inra.fr

M. Bourotte
e-mail: bourotte@avignon.inra.fr

dry conditions correspond to censored values below a given threshold. Positive rainfall are generated by a transform of the Gaussian value above the threshold. This approach is parsimonious because a single latent variable models simultaneously the occurrence of rainfall and its intensity. To transform the Gaussian values above threshold, Kleiber et al. (2013) used a Gamma density. Thompson et al. (2007) and Ailliot et al. (2009) used a power function, but Allcroft and Glasbey (2003) reported that a power transformation is not adequate to achieve normality. They use a quadratic function of the power transformed rainfall. We found that this model is adequate for low and moderate rainfall, but not quite adequate for the most extreme amounts. In Lennartsson et al. (2008), a generalized Pareto distribution (GPD) modeled heavy rainfall above a high level. In Furrer and Katz (2007), a stretched exponential distribution was used as an alternative to the GPD. These models are promising, but they include many parameters.

Our ultimate goal is the disaggregation of daily precipitations into hourly values. Statistical models and their associated simulation algorithms, which constitute precipitation stochastic generators, are clearly well suited to the stochastic disaggregation of precipitation data. Hansen and Ines (2005) disaggregated monthly rainfall into daily values which were modeled using Markov chains for occurrence and mixture of exponential distribution for rainfall amounts. Bürger et al. (2014) proposed to include a temperature dependence in the multiplicative cascade model of (Olsson 1998) for reproducing the Clausius–Clapeyron relation between heavy short-term precipitation and temperature. Hasan and Kunn (2010) used a simple Poisson-gamma model for modeling rainfall occurrence and amount simultaneously. Allcroft and Glasbey (2003) proposed to disaggregate spatio-temporal daily rainfall data with a Gaussian censored latent variable.

Here, we seek a model that can fit the data at these different time scales and for which adequate scaling laws from one time scale to the other can be found. To this end, censored latent models seem the most adequate. We will propose a power-exponential function for transforming the Gaussian values into rainfall amounts. It will be shown that this transformation provides a very good fit from the hourly to the daily time scale, including for the highest values. A striking feature of this model is that we found simple empirical relationships between the parameters estimated at different time scales. We illustrate our model and the disaggregation method on precipitation data collected in two stations located in France: Grignon (about 30 km West of Paris; records from 1996 to 2011) and Chartres (about 80 km South-West of Paris; records from 2001 to 2010).

These two stations are located in the main wheat production region in France. They enjoy a typical Western Europe temperate climate, characterized with a fair amount of precipitation during all seasons. We restrict the

illustration on spring data, collected from March to May. The model is introduced in Sect. 2. Parameter estimation, fitting procedures and scaling relationships are given in Sect. 3. Disaggregation algorithm and results are shown in Sect. 4. Some discussion is in Sect. 5.

2 The model

Let $\{Z_i\}$, $i \in 1, \dots, N$ denote a random process modeling precipitation data, denoted $\{z_i\}$, $i \in 1, \dots, N$. Precipitation data are measured at discrete time intervals, $T, 2T, \dots, NT$, where T denotes the time interval between two consecutive measures. It is thus also the time interval during which precipitation is accumulated in the gauge. If $T = 1\text{h}$, the data represent hourly measurements. Daily measurements correspond to $T = 24\text{h} = 1\text{ day}$. Obviously, each daily rainfall amount must be equal to the sum of the corresponding 24 consecutive hourly measurements. We consider a discrete latent $(0, 1)$ Gaussian process $\{Y_i\}$, $i \in 1, \dots, N$ defined at the same discrete time intervals, with temporal correlation function $c(\cdot)$. The model is fully characterized by two functions: (1) the function that transforms the Gaussian values into precipitations and (2) the correlation function.

2.1 The transform function

We will assume that the transform function is of the form $\psi(y)I(y \geq y_0)$ where $I(y \geq y_0)$ is the indicator function equal to 1 if $y \geq y_0$ and equal to 0 otherwise, and where ψ is a strictly increasing function: $Z_i = \psi(Y_i; \theta)$ if $Y_i > y_0$, and $Z_i = 0$ otherwise, for $i \in 1, \dots, N$. θ is the vector of parameters of ψ . In Ailliot et al. (2009) the transform is a power function, thus ensuring a 1-1 mapping between Y_i and Z_i . In Allcroft and Glasbey (2003), a quadratic form of the power transform is chosen.

We shall take a different route and we will consider a power-exponential transformation of the censored Gaussian values. It was found to offer a very good fit to the data at all time scale considered, see Fig. 1. In particular, there is a fairly good accordance for extreme values. Let us denote $\phi(y)$ the probability density function (pdf) of a $(0, 1)$ Gaussian random variable and $\Phi(y)$ its cumulative probability function (cdf). We shall further denote the complementary cdf $\bar{\Phi}(y) = 1 - \Phi(y)$. The model is thus the following

$$Z = \begin{cases} 0, & Y \leq y_0 \\ z_m + b(e^{a[Y-y_0]^c} - 1), & Y > y_0, \end{cases} \quad (1)$$

where z_m is the resolution of the rain gauge, i.e. it is the minimum quantity a rain gauge is able to measure. typically $z_m = 0.5\text{ mm}$. The parameter y_0 , not included in θ , plays a special role in Eq. (1) because it is directly related to the frequency of dry intervals:

$$P(\text{dry interval}) = P(Z = 0) = P(Y \leq y_0) = \Phi(y_0). \quad (2)$$

The threshold depends on the time scale of the time series. Since dry days are less frequent than dry hours, the threshold y_0 corresponding to daily values will be lower than the threshold corresponding to hourly values.

In the special case $c = 1$, closed form expressions can be derived for the first two moments of rainfall, given that the day is not dry:

$$\begin{aligned} E[Z \mid Z > 0] &= z_m + b(e^{-ay_0} E[e^{aY} \mid Y > y_0] - 1) \\ &= z_m + b \left(e^{a^2/2 - ay_0} \frac{\bar{\Phi}(y_0 - a)}{\bar{\Phi}(y_0)} - 1 \right), \end{aligned} \quad (3)$$

and

$$\begin{aligned} \text{Var}(Z \mid Z > 0) &= b^2 \text{Var}(e^{a[Y - y_0]} \mid Y > y_0) \\ &= b^2 e^{a^2 - 2ay_0} \left[e^{a^2} \frac{\bar{\Phi}(y_0 - 2a)}{\bar{\Phi}(y_0)} - \left(\frac{\bar{\Phi}(y_0 - a)}{\bar{\Phi}(y_0)} \right)^2 \right]. \end{aligned} \quad (4)$$

Equations (3) and (4) were used in Allard (2012) to derive a method of moments estimator for the parameters a and b . The variogram of $Y(\cdot)$ conditional on $Y(\cdot) > y_0$ can also be computed when $c = 1$. Recall that $c(\tau)$ is the correlation function of the latent $(0, 1)$ Gaussian process. Then, from Tallis (1961), technical but otherwise straightforward computations lead to

$$\begin{aligned} \gamma_{Y|Y \geq y_0}(\tau) &= \gamma(\tau) + \frac{\gamma(\tau)^2}{\bar{\Phi}_2(y_0, y_0; c(\tau))} \{ y_0 f(y_0) \bar{\Phi}(y_0^*) \\ &\quad - [2 - \gamma(\tau)] f_2(y_0, y_0; c(\tau)) \}, \end{aligned} \quad (5)$$

with $\gamma(\tau) = 1 - c(\tau)$, $y_0^* = y_0 \sqrt{(1 - c(\tau))/(1 + c(\tau))}$, and ϕ_2 , and $\bar{\Phi}_2$ being the pdf and the complementary cdf of a $(0, 1)$ bivariate Gaussian vector with correlation $c(\tau)$.

2.2 The correlation function

We will consider several models of temporal correlation, with the requirement that parameters should be easy to identify, and that we should be able to establish relationships between parameters at different time scale. We will thus stick to well established models with a low number of parameters. Our first model will be a Matérn correlation function

$$c_1(\tau) = \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{|\tau|}{r} \right)^\nu \mathcal{K}_\nu \left(\frac{|\tau|}{r} \right), \quad \tau \in \mathbb{R},$$

where $\mathcal{K}_\nu$ is the Bessel function of second kind, with smoothness parameter ν and range parameter r .

In temperate regions, the regularity of the rain process, and hence that of the latent process, is close to continuity with absence of differentiability. We thus expect ν to be less than or close to 0.5, the value that corresponds to an exponential covariance function. We also expect the rain

process to exhibit different time ranges, the shorter one corresponding to typical rain events within rainy conditions, while the longer one would correspond to the succession of rainy conditions. Our second model is thus a weighted sum of two exponential models:

$$c_2(\tau) = w \exp(-|\tau|/r_1) + (1 - w) \exp(-|\tau|/r_2), \quad \tau \in \mathbb{R},$$

where r_1 and r_2 are respectively the short and long range parameters and $w \in [0, 1]$. Note however that this work being focused on the disaggregation of daily data, we are mainly concerned with time scales shorter than 24 h. It is therefore possible that a single range is enough. Our third model is thus a simple exponential model:

$$c_3(\tau) = \exp(-|\tau|/r), \quad \tau \in \mathbb{R}.$$

3 Estimation of the parameters

The threshold y_0 is estimated by simply inverting Eq. (2): $\hat{y}_0 = \Phi^{-1}(p)$, where p is the proportion of dry measurements. The other parameters are estimated with a two stage procedure. The parameters of the transform function are first estimated with a marginal likelihood procedure, i.e. by considering the rainfall values as independent. Precipitation amounts are then transformed into Gaussian values by inverting Eq. (1). In the second step, the parameters of the correlation function are estimated using a composite likelihood approach. It was found that ignoring the temporal dependence in the first step did not change the estimates by more than 0.1 % of their values while allowing a faster and more accurate convergence at the second step, in particular for the more complex models c_1 and c_2 .

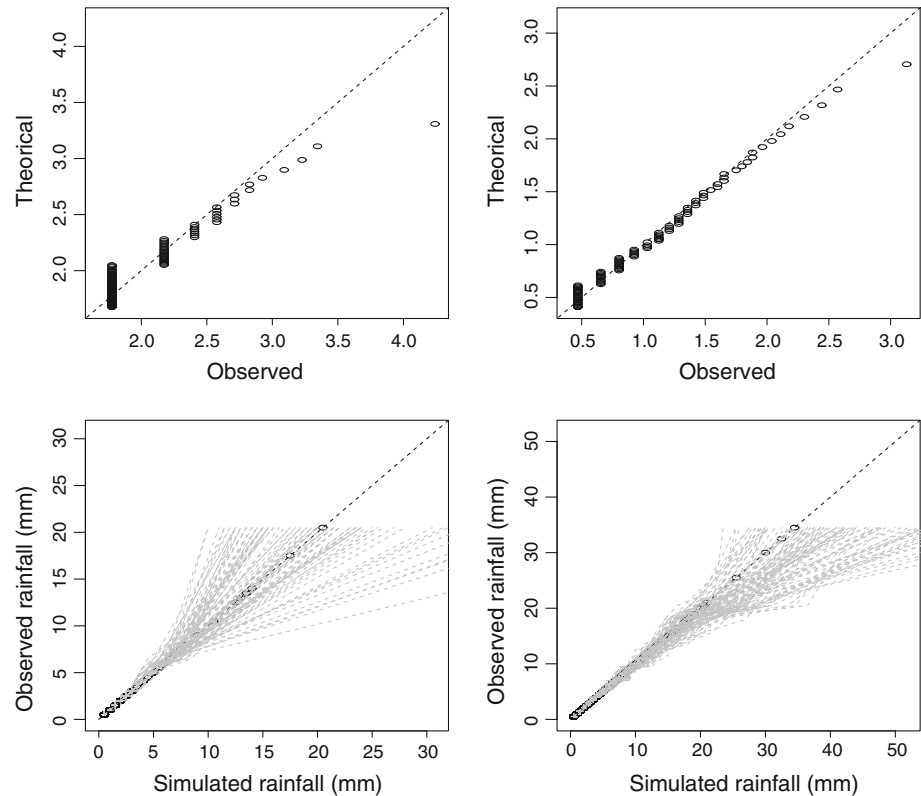
3.1 Estimation of the transform with marginal likelihood

For a given interval T and a given discretization z_m , the parameters $\theta = (a, b, c)$ of the model in Eq. (1) are estimated by simple maximum likelihood. The log-likelihood

$$\begin{aligned} l(\theta) &= \sum_{z_i > z_m^*} \left\{ -\frac{\psi_\theta^{-1}(z_i)^2}{2} - \frac{1}{c} \log a - \log c - \log(z_i - z_m^* + b) \right. \\ &\quad \left. + \left(\frac{1}{c} - 1 \right) \log \left[\log \left(\frac{z_i - z_m^*}{b} + 1 \right) \right] \right\}, \end{aligned} \quad (6)$$

with $\psi_\theta^{-1}(z) = \{a^{-1} \log[b^{-1}(z - z_m^*) + 1]\}^{\frac{1}{c}} + \hat{y}_0$, is numerically maximized by calling the function `dfoptim` in R. For very low precipitations, the discretization effect is proportionally very important (see top plots in Fig. 1). Some values recorded at 0 correspond in reality to precipitations less than or equal to z_m . Conversely, a recorded value equal to z_m can correspond to successive rainfalls less

Fig. 1 Top *Q-Q* plots of spring rainfall data at Grignon; *X-axis* Gaussian quantiles corresponding to the empirical distribution; *Y-axis* Gaussian values as given by Eq. (1). Bottom observed versus simulated *Q-Q* plots of 99 series of data simulated according to the fitted model. Open circles observed values. Grey lines simulated series. Left panels hourly rainfall. Right panels daily rainfall



than or equal to z_m , but whose sum is larger than z_m . The value $z_m^* \leq z_m$ is adjusted to account for the discretization effect. Rain gauges record cumulative amounts that are multiples of z_m . It was found that the best fits were obtained for $z_m^* = 0.375$. From now on, we will set z_m^* to this value.

For the spring series in Grignon, the estimates are $\hat{y}_0 = 1.68$ and $\hat{\theta} = (3.35, 0.075, 0.47)$ for hourly values and $\hat{y}_0 = 0.41$ and $\hat{\theta} = (5.15, 0.026, 0.33)$ for daily values. Note that the parameter c decreases as the time interval increases, as a consequence of the more extreme behavior of daily data resulting from the aggregation of time correlated hourly values. The top plots in Fig. 1 show the *Q-Q*-plots of the Gaussian values obtained for the spring series in Grignon on hourly values (left panel) and daily values (right panel). The agreement is very good, up to the highest value, which lies below the diagonal. In order to validate these fits and to assess the significance of this departure to the diagonal, an ensemble of 99 series of same length was simulated according to the fitted models. The corresponding *Q-Q*-plots are depicted in the lower row of Fig. 1. The observed

values lie on the first diagonal (open circles). The gray lines correspond to the *Q-Q* plots of the simulated series. On both panels, the first diagonal lies in the envelope of simulated lines, which demonstrates that the departure of the highest value from the first diagonal observed on the top plots are within the range of the statistical fluctuations. In addition, considering that our application is oriented towards the simulation of hourly rainfalls conditional on daily ones, these slight departures are not of great concern.

3.2 Estimation of the correlation function

As shown in Allcroft and Glasbey (2003), the correlation coefficients between pairs $\{Z(t), Z(t + \tau)\}$, with $\tau \in \{T, 2T, \dots\}$, can be estimated by maximizing the pairwise log-likelihood $\ell_\tau(z_1, \dots, z_n; \theta, \rho_\tau) = \sum_{i=1}^{n-\tau} \ell_{ij}(z_i, z_{i+\tau}; \theta, \rho_\tau)$, where ρ_τ is the correlation coefficient between Y_i and $Y_{i+\tau}$ and where $\ell_{ij}(z_i, z_j; \theta, \rho_\tau)$ takes one of the three forms, depending on whether neither, one or both measures are dry:

$$\ell_{ij}(\theta; \rho_\tau) = \begin{cases} \log \Phi_2(y_0, y_0; \rho_\tau) & \text{if } z_i = z_j = 0 \\ \log\{\Phi(y_0) - \Phi_2(y_0, y_0; \rho_\tau)\} + \log\{\phi(\psi^{-1}(z_j; \theta))\} & \text{if } z_i = 0 \text{ and } z_j > 0 \\ \log\{\phi_2(\psi^{-1}(z_j; \theta), \psi^{-1}(z_j; \theta); \rho_\tau)\} & \text{otherwise.} \end{cases} \quad (7)$$

Estimating independently all correlation coefficients ρ_τ by maximizing independently the pairwise likelihood for each τ would not automatically lead to a valid correlation function $c(\cdot)$. We thus estimate directly the parameters of the correlation models presented above using a likelihood approach. A full likelihood approach would necessitate heavy computations of multiple integrals in order to take into account all possible successions of dry and wet measurements. We will instead use the weighted pairwise log-likelihood approach presented in Bevilacqua et al. (2012), which is a specific case of the composite likelihood approach (Lindsay 1988). The Weighted Pairwise Log-likelihood (WPL) for the parameters of the correlation function is:

$$\text{WPL}(\eta) = \sum_{i=1}^N \sum_{j=i+1}^N w_{ij} \ell_{ij}(z_i, z_j, \hat{\theta}; \eta) \quad (8)$$

where the values w_{ij} are non-negative weights and η is the vector of parameters of the correlation model. As compared to Eq. (7), there is a slight change of notation, since now the correlation coefficient ρ is obtained from a correlation function with parameters η . There are many possible choices for the weights. The simplest case, which will be chosen here, is when $w_{ij} = 1$ if $|T_i - T_j| \leq \Delta$ and $w_{ij} = 0$ otherwise, in which Δ controls both the efficiency of the estimation and the computational efficiency. We will follow Bevilacqua et al. (2012) for selecting the adequate value Δ by seeking the value

$$\hat{\Delta} = \arg \min_{\Delta \in \mathbb{N}^*} \text{tr}(\mathbf{G}_\Delta^{-1}(\eta_{\text{WLS}}); \Delta). \quad (9)$$

The matrix $\mathbf{G}_\Delta$ is the Godambe information matrix associated to the WPL in Eq. (8):

$$\mathbf{G}_\Delta(\eta) = \mathbf{H}_\Delta(\eta) \mathbf{J}_\Delta^{-1}(\eta) \mathbf{H}_\Delta(\eta)', \quad (10)$$

where

$$\mathbf{H}_\Delta(\eta) = -\text{E} [\text{WPL}(\eta)^{(2)}] \text{ and}$$

$$\mathbf{J}_\Delta(\eta) = \text{E} [\text{WPL}(\eta)^{(1)} \text{WPL}(\eta)^{(1)'}],$$

and $\mathbf{H}_\Delta(\eta)'$ is the transpose matrix of $\mathbf{H}_\Delta(\eta)$. Here, $f^{(1)}$ means the gradient of f (with respect to the parameters in η) and $f^{(2)}$ its Hessian matrix. The inverse of $\mathbf{G}_\Delta(\eta)$ is an approximation of the asymptotic variance of the WPL estimator, $\mathbf{H}_\Delta(\eta)$ is the sensitivity matrix of $\text{WPL}(\eta)$ and $\mathbf{J}_\Delta(\eta)$ is its variability matrix. The minimum of the trace of the Godambe matrix is sought starting with a classical weighted least square estimation η_{WLS} based on the empirical variogram, as advocated in Cressie (1993) and Chilès and Delfiner (2012). We found consistently that the trace of the inverse of the Godambe matrix first decreases rapidly and then reaches a lower floor value usually located

between 12 and 36 h for hourly precipitations and between 4 and 6 days for daily ones. Then the trace increases slightly, but remains very close to the minimum value. Since this indicator is essentially flat around its minimum and because finding the optimal value is computer intensive we decided to set $\Delta = 24$ h for all hourly series and $\Delta = 5$ days for daily ones.

Selecting the correct model of correlation is, as it is the case for all model selection problem, a major issue in statistics. Generally speaking, adding parameters to a model leads to better fitting, and thus to increased values of the log-likelihood. But doing so may lead to over-fitting. Bevilacqua et al. (2012), following Varin and Vidoni (2005), proposed to use a Composite Likelihood Information Criterion (CLIC) for selecting the correlation model. The CLIC penalizes the log-likelihood at the maximum with a negative term, equal to $-2\text{tr}(\hat{\mathbf{J}}_\Delta \hat{\mathbf{H}}_\Delta^{-1})$, which increases with the dimension of the matrices, i.e. with the number of parameters in η . On spatial and spatio-temporal simulations, Bevilacqua et al. (2012) showed that CLIC identified correctly the true model among three in a vast majority of cases, from 82 to 96 % depending on the model chosen for the simulations, when the number of temporal repetition is equal to 200. One drawback of the CLIC is that the penalization term does not depend on the number of data at hand. The Bayesian Information Criterion (BIC) introduced in Schwartz (1978) is a selection criterion that accounts for the number of data used in the estimation procedure. The penalization term is $-\# \eta \cdot \log M$ where $\# \eta$ is the number of parameters in η and M is the number of repetitions. Since BIC was proposed in situations where the data are independent, we must adjust the number of data in order to account for temporal dependency. For doing this, we will use the integral range Lantuéjoul (2002) defined as the integral of the correlation function

$$A(\eta) = \int_{\mathbf{R}} c(\tau; \eta) d\tau \simeq \sum_{k=-\infty}^{\infty} c(kT; \eta). \quad (11)$$

Then, it can be shown that the equivalent number of independent data is equal to $M = N/A(\hat{\eta})$. Finally, we obtain $\text{BIC} = 2 \text{WPL}(\hat{\eta}) - \# \eta \cdot \log M$. Table 1 shows the estimated parameters and the selection criteria for hourly and daily rainfalls on the Grignon spring series. For hourly values, the most adequate model is the double exponential according to both selection criteria. Note that under this model, the correlation at 24 and 48 h are respectively equal to 0.17 and 0.03. For daily rainfall, the Matérn and the double exponential models yield very similar scores. The double exponential is preferred by CLIC (by 1 unit of loglikelihood), whereas the Matérn model is preferred by BIC (by 7 units of loglikelihood).

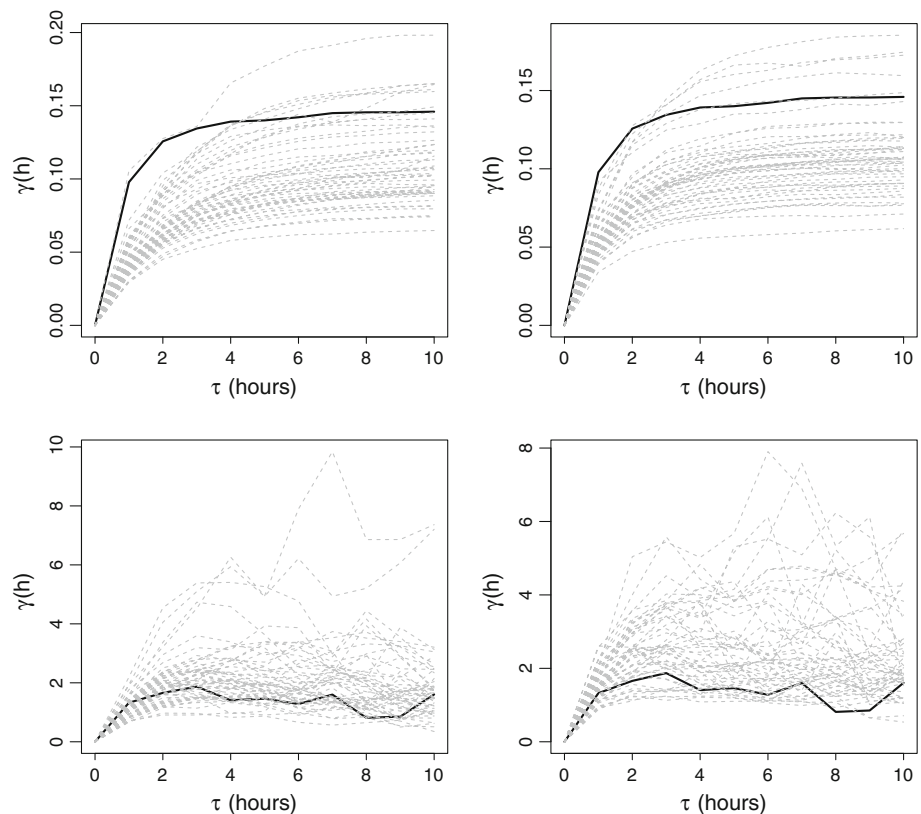
In order to further validate the estimation procedure, we simulated 49 series of realizations of hourly values of same

Table 1 Estimates and selection criteria for the Grignon spring rainfalls

Model	$\hat{r}_1$	$\hat{r}_2$	$\hat{w}$	$\hat{v}$	$\hat{A}$	WPL ($\hat{\eta}$)	CLIC	BIC
Hourly rainfall								
Matérn	7.6	–	–	0.28	10.1	–251,601	–503,218	–503,233
Two Exp.	2.5	15.7	0.60	–	15.6	–251,574	–503,173	–503,192
Single Exp.	5.4	–	–	–	10.8	–251,724	–503,457	–503,463
Daily rainfall								
Matérn	2.7	–	–	0.17	2.7	–10,971	–21,942	–21,966
Two Exp.	0.7	4.7	0.63	–	4.5	–10,970	–21,941	–21,973
Single Exp.	1.6	–	–	–	3.3	–10,983	–21,966	–21,977

Bold value indicates the lowest value for the criterion CLIC, resp. BIC

Fig. 2 Top panels variograms of hourly rainfall amounts of the Grignon spring series (solid black lines) and 49 simulated series with fitted models (dashed grey lines); left panel single exponential model; right panel double exponential model; parameters are given in Table 1. Bottom panels same, computed on hourly non-null rainfall amounts



length as the original series under two models: a single exponential model and the double exponential model selected by CLIC. We then computed the variograms of Z and $Z | Z \geq 0$ for the simulated series and for the genuine rainfall measurements. Results are shown in Fig. 2. The left panels correspond to the series simulated with the single exponential model, which is unable to reproduce the very short correlations. A double exponential model is able to better account for short and long correlations at the same time. Simulated variograms of Z seem less variable than that of the measured series, while simulated variograms of $Z | Z \geq 0$ tend to be slightly more variable than the variogram computed on the positive rainfall amounts.

3.3 Scaling relationships

When hourly data are available, the data set can be transformed at different time scales. For example, bi-hourly data sets can be generated by adding two consecutive hourly values. The same procedure can be easily applied for other time intervals, thus generating data sets synthetically measured at the following time intervals $T = \{2, 3, 6, 12\}$ h. Figure 3 illustrates the variations of the corresponding estimated parameters y_0, a, b, c and r , as a function of T or $\log T$, for the spring series in Grignon and Chartres. As expected, the threshold y_0 decreases as T increases, since the proportion of dry days is lower than the proportion of dry hours. As shown in Fig. 3, we found simple empirical relationships

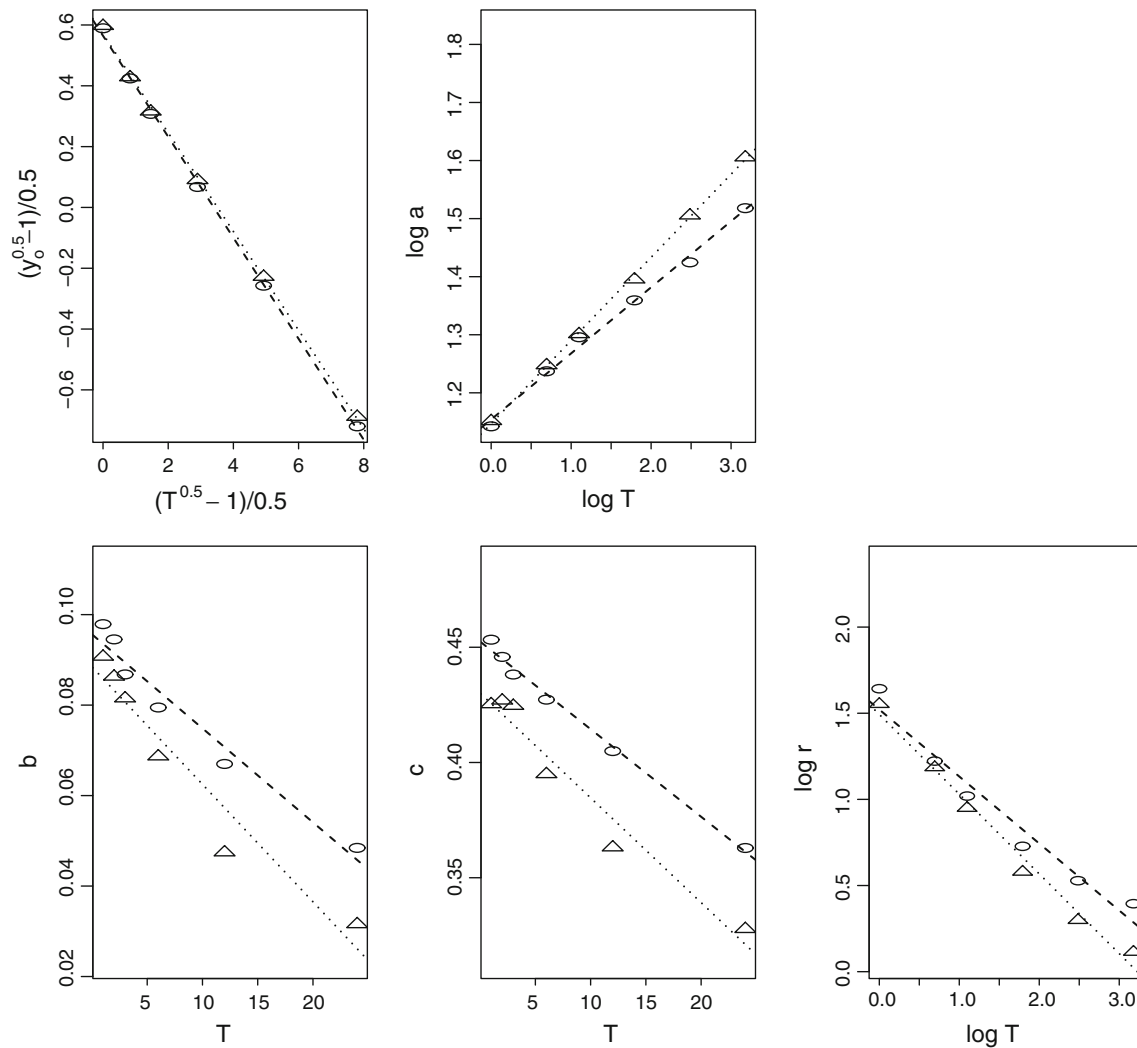


Fig. 3 Empirical relationships between the parameters y_0 , a , b , c and r of the model (1) with respect to the time lag T (in hours). Circles Grignon. Triangles Chartres. Dashed line fitted linear regressions

between the parameters and T . Similar results were consistently observed on other sites analyzed and for all seasons.

4 Disaggregation of daily values

4.1 Algorithm

In order to respect the time arrow, daily values, (z_d) , $d = 1, \dots, N$ are disaggregated into hourly values sequentially. Let $\mathbf{z}_d = (z_{d,1}, \dots, z_{d,24})$ denote the vector of 24 hourly rainfall values of day d and let denote $\mathbf{y}_d$ the corresponding Gaussian values, according to model (1). For day 1, the 24 hourly Gaussian values are simulated conditional on the sum of their transforms being equal to z_1 . Then, from day 2 to day N , hourly rainfall amounts are simulated each day d conditionally on the previous ones. The distribution of $\mathbf{y}_d$

conditional on all previous simulated values can be approximated by a less demanding conditioning on the hourly values of the few previous days concatenated in a vector $\mathbf{y}_p = (\mathbf{y}_{d-1}, \dots, \mathbf{y}_{d-k})$, where k is a small integer. Several values were tested and we found that in our application setting $k = 1$ was sufficient. The distribution $\mathbf{y}_d | \mathbf{y}_p$ is multivariate normal:

$$\mathbf{y}_d | \mathbf{y}_p \sim \text{MVN}(\mathbf{R}_{dp} \mathbf{R}_{pp}^{-1} \mathbf{y}_p, \mathbf{R}_{dd} - \mathbf{R}_{dp} \mathbf{R}_{pp}^{-1} \mathbf{R}_{pd}), \quad (12)$$

where $\mathbf{R}_{fg}$ is the correlation matrix derived from the estimated correlation function computed between vectors $\mathbf{y}_f$ and $\mathbf{y}_g$, with $f, g \in \{d, p\}$. We need to ensure that the total simulated rainfall each day is close to the measured one. The transformation (1) being non linear, there is no straightforward way of simulating these Gaussian values conditional on the sum of their transform being equal to a given value. The simplest algorithm is a rejection

technique: we simply generate vectors from the conditional distribution (12) and accept the proposal if

$$z_d = \sum_{i=1}^{24} z_{d,i} = \sum_{i=1}^{24} \psi(y_{d,i}) I(y_{d,i} \geq \hat{y}_0)$$

is equal to the measured value, up to a pre-specified tolerance ϵ . On dry days, i.e. when $z_d = 0$, the condition becomes $y_{d,i} \leq y_0$, for $i = 1, \dots, 24$. In other words, a truncated Gaussian vector must be simulated on the orthant $\mathcal{O}_{24} = \otimes_{i=1}^{24} (-\infty, y_0)$ with probability density proportional to that of the multivariate Gaussian vector $\mathbf{y}_d | \mathbf{y}_p$ in Eq. (12). The implemented simulation algorithm is thus the following.

Algorithm disagg

1. Read the hourly parameters: a, b, c, y_0
2. Read the parameters of the covariance function and compute the matrices $\mathbf{R}_{dd}$, $\mathbf{R}_{dp}$ and $\mathbf{R}_{pp}$ from this covariance function
3. Simulate $\mathbf{y}_1 \sim \text{MVN}(\mathbf{0}, \mathbf{R}_{dd}) | \{ \sum_{i=1}^{24} \psi(y_{1,i}) I(y_{1,i} \geq \hat{y}_0) = z_1 \}$
4. For $d = 2, \dots, N$:

Until acceptance:

1. Simulate $\mathbf{y}_d \sim \text{MVN}(\mathbf{R}_{dp} \mathbf{R}_{pp}^{-1} \mathbf{y}_p, \mathbf{R}_{dd} - \mathbf{R}_{dp} \mathbf{R}_{pp}^{-1} \mathbf{R}_{pd})$
2. If $|\sum_{i=1}^{24} \psi(y_{d,i}) I(y_{d,i} \geq \hat{y}_0) - z_d| \leq \epsilon$ accept $\mathbf{y}_d$; otherwise reject

The acceptance rate can sometimes be very low (as low as 10^{-6}) but it is compensated by fast computations. To accelerate the algorithm, a slightly more complex algorithm can be implemented. First, parameters at intermediate time intervals such as $T = 12, 6$ or 3 h are estimated, either directly or by using relationships illustrated in Fig. 3. The disaggregation is then performed at each time scale in turn, from the coarsest to the finest, by applying the above technique. From one scale to the next, the length of the vectors to be simulated is dramatically reduced (length 2 or 3). At a given time scale, the earlier vectors must always be simulated before the later ones. Conditioning is on all vectors previously simulated (within the current day and the k previous days) at the same time scale. By doing so, the acceptance rate is multiplied by orders of magnitude and the simulation is accelerated.

4.2 Results

As part of the research project CLIMATOR (Brisson and Levraut 2010), this model was used for disaggregating daily precipitations into hourly precipitations for both measured values (in the past) and large scale climatic model outputs (in the future). These disaggregated

precipitations were used as input values for agricultural and plant models. We were faced with two situations, one being much easier than the other. In the easy situation, some hourly values were available, because they had been measured during some period in the past. In this case, all we had to do was to estimate hourly parameters from the hourly time series, and use these estimates for disaggregating daily rainfalls when hourly ones were not measured.

In the more difficult situation, there were no hourly measurements available. In order to obtain hourly parameters from the daily ones, we used “analog” hourly rainfall time series which were measured either at a nearby station, or at a station considered to belong to a very similar climate. To illustrate this case, daily rainfall at Chartres are disaggregated pretending that hourly rainfall are not available. On the Grignon series, parameters were estimated at different time intervals with an exponential covariance function, i.e. according to model $c_3(\tau)$ for $T = 1, 2, 3, 6, 12, 24$ hours. The regression lines shown in Fig. 3 were then fitted. On this Figure, the genuine parameters in Chartres, along with the corresponding regression lines are shown. It can be seen that the regression lines at the two stations are very close.

Hourly parameters at the Chartres station are obtained by extrapolating the daily parameters in Chartres, $(\hat{y}_0^d, \hat{a}^d, \hat{b}^d, \hat{c}^d, \hat{r}^d)$, with the slopes of the regression lines computed at Grignon. We obtained the following predicted hourly parameters $(\hat{y}_0^h, \hat{a}^h, \hat{b}^h, \hat{c}^h, \hat{r}^h) = (1.70, 3.47, 0.08, 0.42, 3.47)$, which are to be compared to the parameters directly estimated on the hourly precipitation: $(\hat{y}_0^h, \hat{a}^h, \hat{b}^h, \hat{c}^h, \hat{r}^h) = (1.69, 3.16, 0.09, 0.43, 4.72)$.

QQ plots and wet spells of an ensemble of 99 stochastic disaggregations of the spring daily data are represented in Fig. 4. Simulated values have been gridded to the same 0.5 mm scale for comparison with measured data. The resulting QQ plots behave as step functions with fixed increments for low rainfall amounts. Nonetheless, observed data are within the envelope of the 99 simulated series. On this example, the fit is better with the predicted parameters (left panel) than with those directly estimated of the hourly data (right panel). This is not always the case, and in general the envelope contains the diagonal. The distribution of the length of rain events is pretty well simulated, except for a slight underestimation of very short rainfalls. Overall, the distribution and the length of the rain events are well respected. Note that neither on the observed data nor on the simulated series there are rain events longer than 18 h.

5 Summary and discussion

A method for the stochastic disaggregation of daily rainfall values into hourly ones is presented. It has been routinely

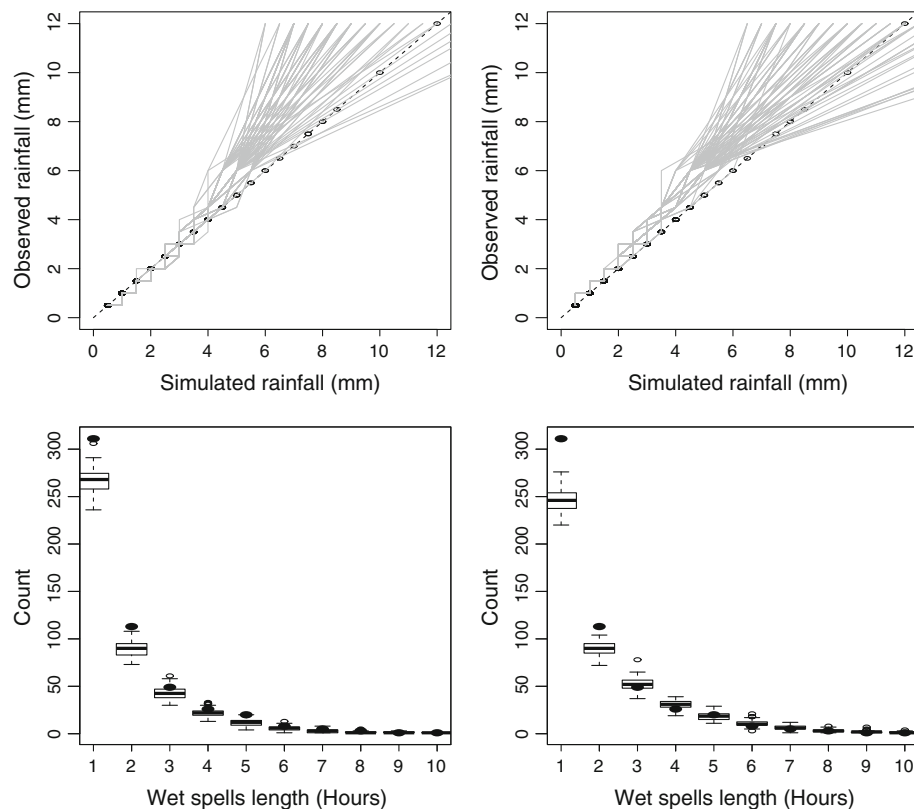


Fig. 4 Top QQ plot of the 99 disaggregated precipitations versus observed hourly precipitations, spring precipitations at the Chartres station. Open circles observed values. Grey lines simulated series. Bottom distribution of the wet spells. Black circles observed series.

Boxplot on the 99 disaggregated hourly series. Left predicted parameters obtained from the regression curves shown in Fig. 3. Right estimated parameters on daily rainfall data

applied for disaggregating daily values at 12 sites representative of the French climatic variability that were part of the CLIMATOR project (Brisson and Levraut 2010). The method relies on a Gaussian latent process, which is transformed into a precipitation process with a power-exponential function. It was found that this transformation function models adequately the body and the tail of rainfall distributions. For space considerations it was only illustrated on the Grignon spring rainfall data, but we found a very good fit on all series analyzed so far. Another consistent finding is that, when analyzing the data at different time scales, we found very good empirical relationships between the parameters and the time scale, see 3. These relationships allow us to predict the parameters of the model at fine time scales that were not measured. Figure 4 illustrates these performances.

We used a sequential algorithm for several reasons. First, from a modeling point of view, it seems natural to respect the time arrow and therefore to simulate one day after the other. Second, taking advantage of this natural order leads to very efficient algorithms. Conditioning each simulated day to all other days (past and future)

necessitates to implement an iterative MCMC algorithm which would be orders of magnitude slower.

This model can be extended in several directions. It is often the case that rainfall is not evenly distributed during the day. To account for this, a straightforward generalization is to allow the threshold and the other parameters to vary within the day. External covariates, if available, could also be included in the statistical model. Even though the CLIC selected the double exponential model as the best model for the auto-correlation of the Gaussian process, the scaling relationships were obtained for the single exponential model. Exploring if such relationships hold for more complex correlation function is another possible extension.

More fundamentally, further work must be undertaken to provide theoretical grounds supporting these empirical findings. In particular, the extreme value properties of this transformation must be assessed and the scaling relationships must be explored under a theoretical point of view. The change of support theory, well-known in geostatistics, is certainly a good starting point for this, see e.g. (Chilès and Delfiner 2012, ch. 6).

As already pointed out in the Introduction, stochastic disaggregation of precipitation is one particular instance of stochastic weather generators. Thanks to the promising results obtained in this work, we believe that latent Gaussian processes with a power-exponential transform are adequate models for precipitation for weather generators.

Acknowledgments We thank Emilio Porcu and Alessandro Fassò for inviting us to join this special issue. Among others, we thank Pierre Ailliot, Valérie Monbet, Philippe Naveau and Carlo Gaetan for fruitful discussions about this work. This work was part of the project CLIMATOR (2007–2010) (http://w3.avignon.inra.fr/projet_climator/) funded by ANR (The French National Research Agency). CLIMATOR was lead by Nadine Brisson who left us in Octobre 2011. We dedicate this work to her memory.

References

- Ailliot P, Allard D, Monbet V, Naveau P Stochastic weather generators : a review of weather type models (submitted)
- Ailliot P, Monbet V (2012) Markov-switching autoregressive models for wind time series. *Environ Model Softw* 30:92–101
- Ailliot P, Thompson C, Thomson P (2009) Space time modeling of precipitation using a hidden Markov model and censored Gaussian distributions. *J R Stat Soc C* 58:405–426
- Allard D (2012) Modelling spatial and spatio-temporal non Gaussian processes. In: Porcu E, Montero JM, Schlather M (eds) *Advances and challenges in space-time modelling of natural events*, Lecture Notes in Statistics 207. Springer, pp 141–164
- Allcroft DJ, Glasbey CA (2003) A latent Gaussian Markov random field model for spatio-temporal rainfall disaggregation. *J R Stat Soc C* 55:1952–2005
- Bevilacqua M, Gaetan C, Mateu J, Porcu E (2012) Estimating space and space-time covariance functions for large data sets : a weighted composite likelihood approach. *J Am Stat Assoc* 107:268–280
- Brisson N, Levraut F (2010) Changement climatique, agriculture et forêt en France: simulations d'impacts sur les principales espèces. *Le Livre Vert du projet CLIMATOR (2007–2010)*. ADEME, Orléans
- Bürger G, Heistermann M, Bronstert A (2014) Towards sub-daily rainfall disaggregation via Clausius–Clapeyron. *J Hydrometeorol* doi:10.1175/JHM-D-13-0161.1
- Caubel J, Launay M, Lannou C, Brisson N (2012) Generic response functions to simulate climate-based processes in models for the development of airborne fungal crop pathogens. *Ecol Model* 242:92–104
- Chen J, Brissette FP, Leconte R (2012) WeaGETS-a Matlab-based daily scale weather generator for generating precipitation and temperature. *Procedia Environ Sci* 13:2222–2235
- Chilès JP, Delfiner P (2012) *Geostatistics: modeling spatial uncertainty*, 2nd edn. Wiley, New York
- Cressie NAC (1993) *Statistics for spatial data*. Wiley, New York
- Flecher C, Naveau P, Allard D, Brisson N (2010) A stochastic daily weather generator for skewed data. *Water Resour Res* 46:W07519
- Furrer EM, Katz RW (2007) Generalized linear modeling approach to stochastic weather generators. *Clim Res* 34:129–144
- Hansen JW, Ines AVM (2005) Stochastic disaggregation of monthly rainfall data for crop simulation studies. *Agric For Meteorol* 131:233–246
- Huber L, Gillespie TJ (1992) Modeling leaf wetness in relation to plant-disease epidemiology. *Annu Rev Phytopathol* 30:553–577
- Hughes JP, Guttorp P, Charles SP (1999) A non-homogeneous hidden Markov model for precipitation occurrence. *Appl Stat* 48:15–30
- Katz RW (1977) Precipitation as a chain-dependant process. *J Appl Meteorol* 16:671–676
- Katz RW, Parlange MB (1995) Generalizations of chain-dependent processes: application to hourly precipitation. *Water Resour Res* 31:1331–1341
- Kleiber W, Katz RW, Rajagopalan B (2013) Daily apatio-temporal precipitation simulation using latent and transformed Gaussian processes. *Water Resour Res* 48:W01523
- Hasan MM, Kunn PK (2010) A simple Poisson-gamma model for modelling rainfall occurrence and amount simultaneously. *Agric For Meteorol* 150:1319–1330
- Lennartsson J, Baxevasi A, Chen D (2008) Modelling precipitation in Sweden using multiple step Markov chains and a composite model. *J Hydrol* 363:42–59
- Lantuéjoul C (2002) *Geostatistical simulations*. Springer, Berlin
- Lindsay B (1988) Composite likelihood methods. *Contemp Math* 80:221–239
- Olsson J (1998) Evaluation of a scaling cascade model for temporal rainfall disaggregation. *Hydrol Earth Syst Sci* 2:19–30
- Racsko P, Szeidl L, Semenov M (1991) A serial approach to local stochastic weather models. *Ecol Model* 57:27–41
- Richardson CW (1981) Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resour Res* 17:182–190
- Schwartz G (1978) Estimating the dimension of a model. *Ann Stat* 6:261–464
- Semenov AM, Brooks RJ, Barrow EM, Richardson CW (1998) Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates. *Clim Res* 10:85–107
- Tallis GM (1961) The moment generating function of the truncated multi-normal distribution. *J R Stat Soc B* 23:223–229
- Thompson C, Thomson P, Zheng X (2007) Fitting a multisite rainfall model to New Zealand data. *J Hydrol* 340:25–39
- Varin C, Vidoni P (2005) A note of composite likelihood inference and model selection. *Biometrika* 92:519–528
- Wilks DS (2010) Use of stochastic weather generators for precipitation downscaling. *Wiley Interdiscip Rev: Clim Change* 1:898–907
- Wilks DS, Wilby RL (1999) The weather generation game: a review of stochastic weather models. *Prog Phys Geogr* 23:329–357
- Zheng X, Katz RW (2008) Simulation of spatial dependence in daily rainfall using multisite generators. *Water Resour Res* 44:W09403

Résultats pour le mois de juillet

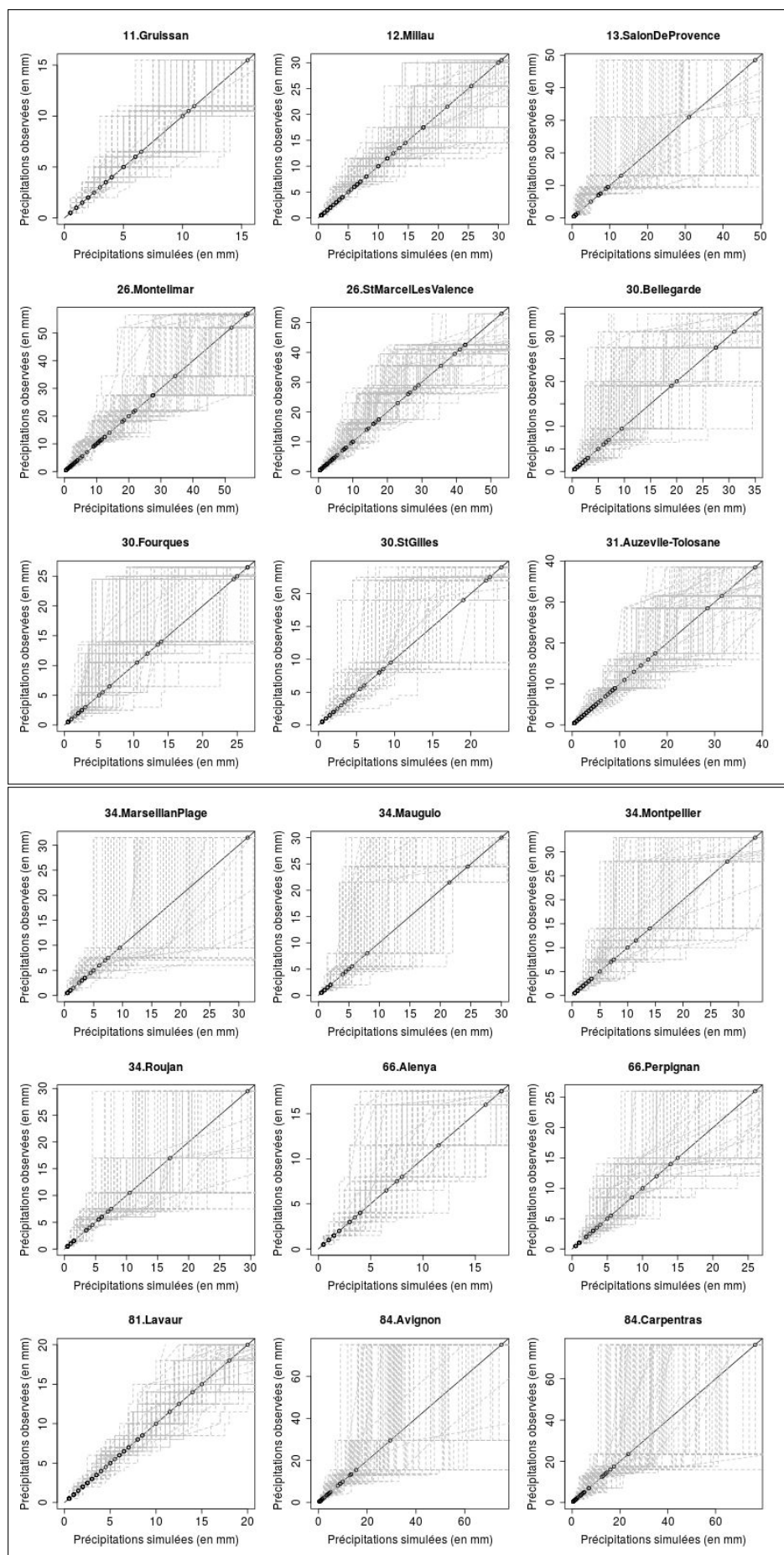


FIGURE A.1 – QQ-plots : précipitations observées vs précipitations simulées (100 simulations) pour le mois de juillet. Les points sur la diagonale correspondent aux vraies valeurs observées.

Station	1	2	3	4	5	6	7	8	9
Paramètre a	7.18	7.02	6.74	7.92	7.33	6.37	5.93	7.08	6.78
Paramètre $b (\times 10^{-2})$	0.68	0.76	1.95	0.64	0.93	2.38	2.54	1.03	0.87
Paramètre $c (\times 10^{-1})$	2.29	3.13	4.57	2.84	3.33	4.44	5.28	3.13	3.12
Station	10	11	12	13	14	15	16	17	18
Paramètre a	5.70	4.95	5.17	5.96	6.31	5.36	6.59	6.50	6.94
Paramètre $b (\times 10^{-2})$	2.30	6.21	3.87	1.97	1.55	3.70	0.84	2.21	1.42
Paramètre $c (\times 10^{-1})$	4.47	5.24	4.99	4.28	3.28	4.20	2.50	3.93	3.35

TABLE A.1 – Estimation par maximum de vraisemblance des paramètres a, b, c de la fonction Ψ pour chacune des stations sur les 11 mois de juillet (2003-2013).