UNIVERSITÉ DU QUÉBEC À MONTRÉAL

CHARACTERISATION OF MULTI-MODEL ENSEMBLES OF CLIMATE-CHANGE PROJECTIONS REGARDING SAMPLING, TREATMENT AND INTERPRETATION

THESIS

PRESENTED

AS PARTIAL REQUIREMENT

FOR PHD DEGREE IN EARTH AND ATMOSPHERIC SCIENCES

BY

MARTIN LEDUC

DECEMBER 2013

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

CARACTÉRISATION DES ENSEMBLES MULTI-MODÈLES DE PROJECTIONS DES CHANGEMENTS CLIMATIQUES AU NIVEAU DE L'ÉCHANTILLONNAGE, DU TRAITEMENT ET DE L'INTERPRÉTATION

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN SCIENCES DE LA TERRE ET DE L'ATMOSPHÈRE

PAR

MARTIN LEDUC

DÉCEMBRE 2013

À ma fille.

REMERCIEMENTS

Je voudrais tout d'abord remercier mon directeur de recherche, René Laprise, pour ses valeureux conseils scientifiques, pour sa grande générosité ainsi que pour la liberté qu'il m'a donnée dans l'élaboration de ce projet de doctorat.

Je tiens aussi à remercier Ramón de Elía qui m'a beaucoup aidé autant d'un point de vue scientifique que personnel ainsi que Léo Séparovic pour les discussions philosophiques et son aide en mathématiques. Aussi, merci à Barbara Casati pour ses commentaires précieux et à Alejandro Di Luca pour sa participation à l'analyse des résultats. Merci au personnel du centre ESCER et du consortium Ouranos de m'avoir permis de travailler dans un environnement stimulant et enrichissant. De plus, je tiens à remercier le centre ESCER ainsi que le programme de *Fonds à l'Accessibilité et à la réussite des Études* (FARE) de l'UQAM pour le financement accordé au cours de mes études.

Merci à Mélissa pour son amour, sa compréhension, pour m'avoir inspiré et accompagné tout au long de mon doctorat. Finalement, merci à mes parents et mes amis pour leur soutien moral via les plaisirs de la vie.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ANOVA	Analysis of Variance
AOGCM	Atmosphere-Ocean General Circulation Model
AR4	IPCC Fourth Assessment Report
CMIP3	Coupled Model Intercomparison Project phase 3
EPP	Ensemble à la physique perturbée
GESA	Gaz à effet de serre et aérosols
GHGA	Greenhouse Gases and Aerosols
GIEC	Groupe d'experts intergouvernemental sur l'évolution du climat
IPCC	Intergovernmental Panel on Climate Change
MCGAO	Modèle de Circulation Générale Couplé Atmosphère-Océan
MMD	CMIP3 Multi-Model Dataset
MME	Multi-Model Ensemble
MMP	Multi-Model Pooling
NARCCAP	The North American Regional Climate Change Assessment Program
PCMDI	Program for Climate Model Diagnosis and Intercomparison
PPE	Perturbed Physics Ensemble
RCM	Regional Climate Model
SDM	Statistical Downscaling Model
SMP	Single-Model Pooling
SRES	Special Report on Emissions Scenarios
TAR	IPCC Third Assessment Report

RÉSUMÉ

Cette thèse traite de diverses difficultés inhérentes à l'analyse d'ensembles multi-modèles de projections de changements climatiques. Ces ensembles, souvent appelés « ensembles d'opportunité », sont formés en fonction de la disponibilité de plusieurs centres de modélisation à l'échelle mondiale à produire un certain nombre de simulations. Les ensembles résultants d'un tel processus ne sont donc pas construits selon un cadre expérimental systématique visant à permettre une analyse optimale, mais plutôt en fonction de facteurs externes émergeant d'un processus d'échantillonnage ouvert.

Dans le premier chapitre de cette thèse, le concept d'un échantillonnage de type « expert » est étudié. Consistant en une présélection d'un certain nombre de simulations à partir de l'ensemble disponible, ce type de processus est généralement utilisé dans le but de réduire la taille d'un ensemble qui ne peut être traité en entier lorsque les ressources sont limitées. Les incertitudes d'échantillonnage reliées au calcul des statistiques de l'ensemble sont calculées en ré-échantillonnant sur un grand nombre de sous-ensembles de simulations. Le processus de sélection est divisé en deux types de choix faits par les experts : le choix des modèles et celui des membres. Il est démontré comment ces incertitudes d'échantillonnage consistent en des manifestations de sources d'incertitudes connues reliées aux projections climatiques, soient la variabilité climatique naturelle et l'écart-type inter-modèle.

Le second chapitre vise à étudier une problématique fondamentale à l'échantillonnage des modèles dans un ensemble d'opportunité. Les modèles de climat n'étant *a priori* pas tout à fait indépendants puisque les scientifiques partagent des connaissances à propos du système climatique et quant à la manière de construire les modèles, aucune métrique pour évaluer cette indépendance ne fait présentement consensus entre les scientifiques. Dans ce chapitre, nous proposons un critère pour détecter un manque d'indépendance entre les projections de changements climatiques. Ce critère est basé sur le fait que deux modèles peuvent mener à des sensibilités climatiques similaires face aux forçages externes, mais un tel consensus devrait être rejeté quand des raisons suffisantes peuvent remettre en cause la notion d'indépendance. Par exemple, lorsque d'importantes similarités structurelles apparaissent entre les modèles ou, dans une moindre mesure, dû à une certaine dépendance institutionnelle.

Dans le troisième chapitre, des pistes de solutions sont suggérées face au problème que les modèles sont généralement représentés dans un ensemble par peu de membres et en nombres souvent inégaux. L'utilisation d'échantillons non-équilibrés peut engendrer certains problèmes, particulièrement en ce qui a trait à l'estimation de la variabilité naturelle dans l'ensemble, celle-ci étant souvent obtenue à partir de l'écart-type inter-membre.

Avant de considérer des méthodes de reconstruction visant à régénérer les simulations jugées manquantes à partir de l'information disponible dans l'ensemble, deux hypothèses se doivent d'être vérifiées. La première s'applique à un ensemble de membres provenant d'un seul modèle et consiste à déterminer si cet ensemble peut être supposé comme étant *ergodique*, c.-à-d. que la variabilité temporelle est à peu près égale à celle qui intervient entre les membres. La seconde hypothèse considère que la variabilité naturelle est simulée de façon égale entre les modèles. Bien que les résultats montrent que la variabilité naturelle diffère de façon importante entre les modèles, l'hypothèse d'ergodicité entre les membres s'avère vraie pour des simulations sans forçages externes. Pour des simulations avec forçages externes, il est démontré comment des conditions de stationnarité peuvent être atteintes par traitement en soustrayant les tendances polynomiales dans les séries temporelles.

Dans le quatrième chapitre sont comparées différentes méthodes pour quantifier la variabilité naturelle à partir d'une combinaison de plusieurs modèles. D'un côté, l'estimé optimal pour cette variabilité serait biaisé vers les modèles avec le plus de membres, tandis qu'un estimé donnant le même poids à tous les modèles serait caractérisé par une plus grande erreur type. Dans ce même chapitre est aussi fourni un exemple d'application de l'hypothèse d'ergodicité, qui permet d'utiliser la variabilité temporelle afin de comparer les signaux de changements climatiques provenant de deux modèles, lorsque ces derniers sont représentés par un seul membre. Cette approche peut être vue comme une alternative devant la méthode plus coûteuse de considérer des expériences supplémentaires, par exemple les simulations de contrôle pour la période préindustrielle disponibles dans l'ensemble CMIP3.

Mots clés : ensemble multi-modèle, échantillon non balancé, variabilité naturelle, incertitude modèle, indépendance des modèles, ergodicité

ABSTRACT

This thesis focuses on inherent issues to the analysis of multi-model ensembles of climate-change projections. Such ensembles, often denoted as "ensembles of opportunity", are formed on the basis of the readiness of several modelling centres around the world to produce simulations. It results in ensembles that are not constructed based on a systematic framework aimed at an optimised analysis but rather on external factors emerging from an open sampling process.

In the first chapter of this thesis, the concept of an expert-based sampling is investigated, consisting in making a pre-selection of a number of simulations from a large ensemble. Such a sampling process is generally used by research centres that cannot afford to handle the entire ensemble due to limited resources of treatment. Sampling uncertainties affecting the statistics of the resulting ensemble are assessed using resampling methods by randomly selecting over several ensembles subsets. The selection process is divided as two types of choices made by the experts : the choice of the models and that of the members. We show how these sampling uncertainties are manifestations of known sources of uncertainty, namely the natural climate variability and the inter-model spread.

The second chapter investigates an issue that is fundamental to the sampling of the models in an ensemble of opportunity. While climate models are not expected to be independent since scientists share knowledge about the climate system and on how to construct models, no robust metric to quantify model independence is commonly accepted among climate scientists. In this chapter, we propose a criterion for detecting possible lack of independence between climate-change projections. This criterion is based on the fact that two models can lead to similar climate sensitivities to external forcings, but such a consensus should be rejected when there are sufficient reasons to believe that it occurs for the wrong reasons, i.e. whether due to important structural similarities between the models or to a lesser extent, to some institutional dependence.

In the third chapter, a workaround to the apparent problem of a small and unequal number of members provided by the models is investigated. Such an imbalance between sample sizes raises issues in the assessment of the natural climate variability when obtained from the inter-member spread. When considering reconstruction methods for regenerating these "missing simulations", two assumptions about the multi-model ensemble have to be investigated. The first one applies to a single model and consists in determining whether an ensemble of members can be assumed as *ergodic*, i.e. that the variability measured in time is approximately equal to the inter-member spread. The second assumption is that the natural variability is equally simulated by the different

models in the ensemble. While the results show that the natural variability largely differs among the models, an ensemble of members can be considered as ergodic when run under stationary conditions. For simulations run under transient forcings, it is shown how stationary conditions can be reached by treatment by removing polynomial trends from the time series.

In the fourth chapter, different methods are compared for assessing the natural variability from a multi-model ensemble. While an optimal estimator of the natural variability would be biased toward the models with larger sample sizes, an unweighted estimate that gives an equal importance to the different models would be affected by larger sampling errors. We also provide an example of application of the ergodic assumption that allows taking advantage of the temporal variability in the simulations in order to compare the climate-change signals provided by two models when both provide a single member. This method can be seen as an alternative to the more expensive way of using supplementary simulations run without external forcings such as the pre-industrial control experiments in the CMIP3 multi-model ensemble.

Keywords : multi-model ensemble, unbalanced framework, natural variability, model uncertainty, model independence, ergodicity

INTRODUCTION

La méthode scientifique requiert que les théories soient validées par l'expérimentation. Toutefois, en science du climat, les chercheurs n'ont pas accès à un laboratoire au sens classique qui permette de vérifier leurs hypothèses. En ce sens, le système climatique terrestre est à la fois laboratoire et sujet d'étude. Considérant que certaines perturbations du système climatique peuvent prendre plusieurs décennies avant que les répercussions puissent être ressenties de manière significative, il serait peu judicieux pour l'Homme d'envisager de perturber son environnement afin d'en évaluer les conséquences.

Les modèles de climat

Les scientifiques du climat doivent donc se tourner vers des expériences effectuées par ordinateur où les équations mathématiques décrivant la physique du système climatique permettent d'en simuler l'évolution. Au cours des dernières décennies, la science du climat a évolué considérablement, et ce, en grande partie grâce à l'augmentation de la puissance de calcul des ordinateurs. Les principaux outils à la portée des scientifiques sont les Modèles de Circulation Générale Couplés Atmosphère-Océan (MCGAO ; Randall et al. 2007), qui tiennent compte des principales composantes du système climatique : l'atmosphère, les océans, la surface terrestre, la glace de mer et la biosphère. Dans ces modèles sont prescrits des forçages dits "externes" comme les émissions de gaz à effet de serre et d'aérosols (GESA) (Nakicenovic et al., 2000). À l'aide des MCGAO contemporains, le climat planétaire peut être simulé sur plusieurs centaines d'années à des résolutions spatiales de l'ordre d'une centaine de kilomètres, et ce, en quelques semaines de calcul sur un superordinateur. Le coût relatif à la production de ces simulations reflète à quel point les modèles de climat sont des programmes informatiques complexes nécessitant une grande puissance de calcul.

Incertitude dans les projections climatiques

En dépit de la grande complexité des modèles de climat, ces derniers ne restent que des approximations du système climatique réel. D'abord par leur nature discrète, ils ont une résolution finie, et donc même certains processus assez bien connus comme la dynamique des fluides se voient alors approximés. De façon similaire, d'autres approximations ont lieu puisque certains processus physiques interviennent à des échelles plus fines que la grille du modèle. Ces processus non résolus par le modèle, par exemple la convection, la micro-physique des nuages ou les transferts radiatifs, se doivent donc d'y être intégrés sous forme de paramétrages (Tompkins, 2002).

Les projections climatiques sont évidemment sujettes à un certain niveau d'incertitude. Cette incertitude peut être séparée en trois composantes, soit la variabilité naturelle du climat, l'incertitude reliée aux approximations utilisées par un modèle et l'incertitude due au choix de scénario de GESA (Hawkins et Sutton, 2011). La variabilité naturelle est une composante fondamentale d'incertitude puisqu'elle reflète le caractère chaotique du système climatique (Lorenz, 1963). Cette source de variabilité est générée à l'intérieur même du système et est souvent considérée comme le niveau minimal de "bruit climatique" en deçà duquel le système ne peut être considéré déterministe. La variabilité naturelle générée par un modèle de climat peut être quantifiée de deux manières différentes. La première consiste à générer une longue simulation (e. g. plusieurs centaines d'années) et d'en évaluer la variabilité temporelle (Deser et al., 2010). La seconde consiste à générer plusieurs réalisations d'un même climat en imposant de petites différences dans les conditions initiales. Par la nature chaotique du système, ces simulations perdront toute mémoire de leurs conditions initiales après une certaine période de temps de mise à l'équilibre (Stouffer et al., 2004; Stouffer, 2004). La variabilité entre ces différentes réalisations est souvent utilisée comme mesure de la variabilité naturelle simulée par un modèle de climat (Sorteberg et Kvamstø, 2006 ; Deser et al. 2010).

L'incertitude modèle est due au fait que les scientifiques ont une connaissance limitée

du système climatique. Autant le choix des processus physiques d'intérêt à inclure dans les modèles que la manière de les transposer sous forme d'équations pouvant être solutionnées par ordinateur peut différer entre les experts. Les modèles sont donc construits différemment, ce qui mène à certaines différences dans leurs projections climatiques. L'incertitude due au scénario est due au fait que l'évolution future des émissions anthropiques de GESA est pratiquement inconnue. Ces émissions dépendent notamment de l'évolution du contexte socio-économique, technologique et politique mondial. Elles sont donc très difficiles à prévoir et cette question dépasse largement le cadre de la problématique reliée à la modélisation du système climatique. Or, l'utilisation de différents scénarios d'émissions dans les simulations climatiques montre clairement l'effet de ces derniers sur l'ampleur et les détails du changement climatique appréhendé (Meehl et al., 2007a), faisant du choix de scénario une source importante d'incertitude dans les projections climatiques.

Les ensembles d'opportunité

Dans le but de quantifier les différentes sources d'incertitude reliées aux projections climatiques, d'imposants ensembles de simulations doivent être utilisés. En mettant à contribution les différents centres de recherche en modélisation climatique de par le monde, ces projets internationaux permettent un certain échantillonnage des différentes sources d'incertitude. Un bon exemple de ce type d'ensemble est la phase 3 du projet d'intercomparaison de modèles couplés (CMIP3; Meehl et al. 2007b). Cet ensemble contient des simulations provenant de plus d'une vingtaine de modèles pour quelques scénarios d'émissions de GESA. La variabilité naturelle y est échantillonnée à l'aide de plusieurs réalisations par expérience, de même que par un certain nombre de simulations de la période préindustrielle où aucun forçage anthropique n'est appliqué.

Le processus d'échantillonnage de l'ensemble CMIP3 reste relativement ouvert en ne posant que certaines conditions minimales aux différents centres pour y participer. Ceci permet entre autres de maximiser le nombre de modèles dans l'ensemble. Ces conditions

minimales peuvent se résumer à utiliser un MCGAO conforme aux règles de l'art pour générer un certain nombre de simulations en fonction d'expériences suggérées, et ce, dans les délais et formats d'archivage requis par le projet. Un tel processus d'échantillonnage engendre un ensemble dont la structure est principalement définie par l'offre en simulations, soit le degré de participation des différentes équipes de recherche en fonction de leurs ressources et intérêts. Au final, l'ensemble sera souvent incomplet, c'est-à-dire que tous les modèles ne sont pas utilisés pour générer toutes les expériences proposées étant donné le coût important relié à la production de telles simulations. Pour les mêmes raisons, l'ensemble a peu de chances d'être équilibré, et donc que les modèles et institutions y sont représentés de façon plutôt inégale selon les trois axes d'incertitude.

Problèmes inhérents aux ensembles multi-modèles

L'échantillonnage des principales sources d'incertitude via ce type d'ensemble pose cependant plusieurs problèmes. D'abord, par sa structure irrégulière, l'analyse d'un ensemble multi-modèle mène à des approximations dans les méthodes statistiques conventionnelles (von Storch et Zwiers, 1999) et possiblement à des biais. Or, ce type de problème n'est pas nouveau, Kendall (1946) ayant déjà mentionné l'importance d'impliquer des mathématiciens lors d'un processus échantillonnage afin de permettre l'application d'une analyse de type exact (i.e. sans approximations), où les biais sont minimisés et les erreurs d'échantillonnage contrôlées. Dans le cas de CMIP3, on peut voir ces problèmes comme un compromis étant donné le processus d'échantillonnage ouvert permettant la maximisation du nombre de modèles dans l'ensemble.

Un exemple d'ensemble où ces problèmes sont considérés lors du processus d'échantillonnage est le projet NARCCAP (*The North American Regional Climate Change Assessment Program* ; Mearns et al. 2009). La structure de l'ensemble y est déterminée à l'avance afin d'en optimiser l'analyse. On notera aussi le projet ENSEMBLES (van der Linden et Mitchell, 2009), qui au même titre que CMIP3, utilise un processus d'échantillonnage basé sur l'offre en simulations, résultant en une structure d'ensemble incom-

plète et non équilibrée. Dans le but d'analyser les différentes composantes d'incertitude reliées à cet ensemble, Déqué et al. (2012) a dû utiliser certaines astuces mathématiques afin de reconstruire les expériences manquantes dans la structure. Un avantage d'une telle approche est d'obtenir un cadre expérimental souhaitable pour l'application d'une méthode d'analyse exacte en évitant les biais lors de l'évaluation des différentes composantes d'incertitude.

Cependant, même dans le cas d'un ensemble multi-modèle dont la structure est complétée et équilibrée selon les différentes expériences suggérées, certains problèmes d'échantillonnage persistent au-delà de ceux strictement reliés la structure même de l'ensemble. Un problème de taille réside dans l'échantillonnage de l'incertitude modèle. Typiquement, l'incertitude modèle est étudiée à l'aide de deux types d'ensemble. Le premier est l'ensemble à la "physique perturbée" (EPP) qui consiste à utiliser un seul modèle sous différentes configurations. Ces configurations sont obtenues en variant certains paramètres du modèle dont la valeur est incertaine (Rowlands et al., 2012). Un modèle pouvant contenir des centaines de paramètres à varier, l'étude de l'incertitude modèle via ce type d'ensemble consiste à explorer un espace avec autant de dimensions, ce qui est hors de portée pour la plupart des groupes de recherche en modélisation. Un effort considérable dans ce domaine est le projet *climateprediction.net* (Stainforth et al., 2005) qui utilise des ressources informatiques distribuées afin de générer un ensemble comptant plusieurs milliers de simulations. Cependant, un EPP reste par définition contraint aux particularités structurelles d'un seul modèle et donc ne révèle qu'une facette de l'incertitude modèle (Tebaldi et Knutti, 2007).

La deuxième manière d'étudier l'incertitude modèle consiste à utiliser un ensemble multi-modèle (e.g. CMIP3). Ce type d'ensemble tient compte des différences structurelles entre les modèles, comme le choix des processus d'intérêt à considérer ou la manière de les représenter sous forme de paramétrages. Un problème important relié à ce type d'ensemble est que les modèles y sont échantillonnés de manière ni aléatoire ni systématique, mais plutôt basée sur la disponibilité des modèles (offre en simulations). L'échantillonnage

multi-modèle explore un espace indéfini qui ne peut être simplement représenté à partir de nombres comme c'est le cas pour l'EPP dont l'espace des paramètres est défini, bien que extrêmement coûteux à explorer (Murphy et al., 2007). Les difficultés reliées à la définition d'un "espace des modèles" reposent sur une problématique d'ordre conceptuelle. Cette difficulté constitue une importante barrière devant toute interprétation probabiliste des résultats de l'ensemble, à moins d'utiliser des hypothèses substantielles (Giorgi et Mearns, 2002; Greene et al. 2006).

Un point important relié à l'échantillonnage des modèles est que plusieurs raisons portent à penser qu'ils ne sont pas tout à fait indépendants l'un de l'autre. En fait, les centres de modélisation partagent des connaissances en ce qui a trait au système climatique et à la manière de construire les modèles, par la participation à des conférences et la publication d'articles spécialisés. De plus, les modèles sont souvent validés et ajustées (via leurs paramètres) en fonction de données climatiques similaires. Un indicateur de ce manque d'indépendance est que les modèles ont en commun certains biais lorsque leurs résultats sont comparés avec le climat observé (Lambert et Boer, 2001; Knutti et al. 2010). En guise de comparaison, des échantillons indépendants devraient statistiquement mener à une annulation des erreurs au fur et à mesure que la taille de l'ensemble est augmentée, ce qui n'est pas le cas pour les MCGAOs contemporains. De plus, Masson et Knutti (2011) ont mis en évidence que des similarités entre les résultats de modèles tendent à apparaître lorsque ces derniers sont développés par des acteurs communs. Aucune métrique ne faisant présentement consensus parmi les scientifiques afin de quantifier le concept d'indépendance (Tebaldi et Knutti, 2007), certaines implications sont très importantes. Par exemple, l'utilisation d'une norme basée sur les similarités des simulations des modèles en guise d'indicateur de confiance dans un résultat donné se voit une idée difficile à défendre sans une confiance de l'indépendance des modèles (Pirtle et al., 2010); les similarités pouvant très bien apparaître pour les mauvaises raisons, par exemple dû à des hypothèses similaires utilisées dans la construction des différents modèles.

Objectifs et plan de la thèse

Cette thèse vise à mettre en lumière plusieurs problématiques fondamentales auxquelles doivent faire face les scientifiques lors de l'analyse d'un ensemble multi-modèle. L'ensemble CMIP3 y est utilisé à titre d'exemple mais ces problématiques se veulent tout aussi applicables à d'autres ensembles comme CMIP5. En particulier, on s'attarde aux deux sources d'incertitude primordiales des projections climatiques, c'est-à-dire la variabilité naturelle et l'incertitude modèle. La thèse est divisée en quatre chapitres qui représentent des articles scientifiques à soumettre à des revues spécialisées.

L'ensemble CMIP3 étant le résultat d'un effort sans précédent de coordination à l'échelle mondiale, il est donc très riche en information mais aussi relativement imposant en termes de volume de données. Une équipe de recherche utilisant ces simulations se limitera souvent à n'utiliser qu'une partie de l'ensemble selon ses capacités de traitement de données et des questions scientifiques à étudier. Ce processus de sélection d'un ensemble est fait par les experts et vise d'abord à réduire la taille de l'ensemble mis à leur disposition tout en minimisant les pertes en information. Dans le premier chapitre, on étudie les erreurs d'échantillonnage issues d'une présélection de simulations quant à leur effet sur le calcul des statistiques de l'ensemble. Le processus d'échantillonnage par les experts y est analysé en fonction d'une sélection faite sur les modèles ainsi que sur les membres disponibles pour chaque modèle. Le cadre expérimental proposé vise entre autres à mieux comprendre les effets d'un ensemble de taille finie en évaluant les erreurs statistiques en fonction de la taille de l'échantillon sélectionné. On y discute notamment les hypothèses fondamentales qui se doivent généralement d'être adoptées lors de l'utilisation de ces ensembles.

Après l'étude du processus de sélection d'un ensemble par les experts, le second chapitre traite de la nature de l'échantillonnage à la base même d'un ensemble multi-modèle. D'abord, on y discute de la participation des centres de recherche et de leur effet sur l'échantillon disponible. Le concept d'indépendance des modèles y est ensuite révisé en

détails selon les travaux déjà abordés. Nous proposons par la suite un cadre expérimental visant à quantifier la notion d'indépendance quant aux consensus et désaccords observés entre les signaux de changements climatiques par rapport au niveau de bruit donné par la variabilité naturelle. On y étudie aussi l'hypothèse souvent évoquée que deux modèles développés par une même institution tendent à donner des résultats avec des caractéristiques similaires. Bien que cette hypothèse ne soit pas toujours vraie, elle reste néanmoins un outil important en vue de filtrer l'ensemble de ses consensus non informatifs, c'est-à-dire dus aux mauvaises raisons. Enfin, on y avance certaines pistes de solution qui devraient être considérées par la communauté scientifique afin de diminuer l'ampleur du problème relié au manque d'indépendance entre les modèles pour les ensembles à venir.

Dans le troisième chapitre, un autre type d'échantillonnage est abordé. On le qualifiera d'échantillonnage synthétique, celui-ci visant à régénérer artificiellement les simulations considérées comme manquantes dans l'ensemble. Ce type d'approche est principalement voué à simplifier l'analyse d'un ensemble incomplet ou non équilibré à l'aide de méthodes peu coûteuses en comparaison avec la production de simulations à l'aide d'un MCGAO. Ce chapitre propose notamment deux types d'approches visant à tirer profit de l'information temporelle disponible dans l'ensemble en vue d'y générer de nouveaux membres. La première technique consiste à utiliser l'information temporelle fournie par un modèle afin de lui générer des membres supplémentaires, tandis que la seconde consiste à utiliser l'information temporelle provenant de tous les modèles de l'ensemble. Ces deux approches sont placées dans un cadre décisionnel afin de déterminer la méthode souhaitable en fonction de l'ensemble utilisé. En particulier, la première méthode évoque le caractère *ergodique* d'un ensemble de membres provenant d'un seul modèle. Cette caractéristique apparaît comme une symétrie entre le temps et les membres ; elle peut être d'une grande utilité pour la reconstruction d'expériences manquantes dans un ensemble multi-modèle.

Le quatrième chapitre se veut une récapitulation des principaux concepts développés

dans cette thèse. On y propose notamment deux exemples d'application. Le premier fait une comparaison entre différentes approches afin de combiner la variabilité naturelle simulée par différents modèles. Le deuxième exemple applique le principe d'ergodicité entre les membres afin d'améliorer la qualité des tests statistiques proposés dans le premier chapitre quant à la caractérisation de l'indépendance entre deux modèles développés par un même groupe de modélisation.

CHAPTER I

ON THE UNCERTAINTY RELATED TO EXPERTS' DECISIONS IN THE SELECTION OF A SUBSET OF SIMULATIONS FROM A LARGE ENSEMBLE OF OPPORTUNITY

ABSTRACT

From the climate modelling point of view, an ensemble of opportunity consists of a group of simulations generated using several models developed by different research centres around the world. Such ensembles are generally formed in a rather open way by allowing research groups to provide an arbitrary number of member simulations generated from one or several versions of their model. While these simulations are used in a wide variety of applications, users often consider only a small part of the entire available ensemble due to limited resources for data handling.

In this chapter, we investigate the concept of the sampling uncertainties emerging from the selection of a subset of simulations from a large ensemble. It is shown how these uncertainties can be constrained by the selection process and the underlying assumption about the nature of the ensemble-related population. Emerging as the lower bound of error in the ensemble statistics, these sampling uncertainties consist in different manifestations of known sources of uncertainty in climate modelling such as the natural variability and the model structural uncertainty.

1.1 Introduction

As a result of different approximations and alternative approaches employed, different coupled Atmosphere-Ocean General Circulation Models (AOGCMs) developed by a number of research teams around the world give different climate sensitivities in response to the same concentration of greenhouse gases and aerosols (GHGA). In order to

interpret these differences and understand their impacts on climate-change projections for the next century, some internationally coordinated efforts have been realised over the last years, aiming at setting up experimental frameworks that allow comparing and combining climate simulations from different models. These large projects are formed in a rather open way, meaning that research centres are generally free to participate by delivering an arbitrary number of simulations. Such experiments allow collecting a relatively large number of simulations, leading to a range of credible climate-change projections that are brought together as a multi-model ensemble of simulations. At this time, the best achieved example of such an application is the Coupled Model Intercomparison Project Phase 3 (CMIP3; Meehl et al. 2007b) while CMIP5 is underway at the time of writing.

A fundamental issue of climate-change modelling resides in the intrinsic nature of multi-model ensembles. Often denoted as "ensemble of opportunity" (e.g. Christensen et al. 2007, Tebaldi and Knutti 2007, Annan and Hargreaves 2010), such ensembles do not imply any random or systematic sampling of the models over the possible population of all modelling approaches. Research centres around the world are free to participate to the coordinated effort towards climate-change assessment, but they do so according to their own computing and human resources constraints. This results in ensembles that sample in some way the model structural differences (or modelling approaches); the spread among simulations is often interpreted as reflecting the uncertainty of climate-change projections, in addition to the uncertainty about the future GHGA emissions pathways. It is worth noting that participating groups provide an arbitrary number of realisations from the same model, usually referred to as "members", which sample the models' natural variability (Sorteberg and Kvamstø, 2006; Deser et al., 2010). Moreover, the rather open method of forming an ensemble allows for a participating group to provide runs from several versions of the same model, that may differ for example by changes of spatial resolution, parameterization packages, or different tuning of some parameters.

While the simulations resulting from these ensembles are often used "as is" in a va-

riety of climate-change assessments, downscaling techniques are increasingly used in the hope of obtaining further regional information. Examples are dynamical downscaling with Regional Climate Models (RCMs; Rummukainen 2010) and statistical downscaling (e.g. Dibike et al. 2008) in order to obtain small-scale details from the coarse-resolution AOGCMs' simulations. Either approaches involve a large amount of data-handling resources, and computing resources in the case of RCM; hence the motivation for considering only a subset of the original AOGCM ensemble. Expert decisions may be involved in selecting such a subset in order to minimise losses of valuable information. One common approach for reducing a large ensemble into a smaller subset is by retaining a single member of each model or version of model (e.g. Bombardi and Carvalho 2011, Peings and Douville 2010, Räisänen et al. 2010) when several are available, thus reducing the size of the ensemble to the number of available models. Such sampling is expected to have a little impact for climate-change projections made over several decades, since at these time scales, the inter-member variability is generally smaller than the inter-model variability (Hawkins and Sutton, 2009). The idea of retaining a single member per model also sustains the democratic idea of "one vote per model" (Knutti 2010) in the assessment of the climate-change signal.

If such "one member per model" reduced dataset is still too large for the handling capability of a user, the ensemble is further reduced by proceeding to the selection of a smaller number of models according to some specific characteristics. An often used criterion is the models' performance in reproducing the present climate (Gleckler et al. 2008) in order to remove from the ensemble the models that are considered less reliable. Another one consists in eliminating the "outliers" whose climate change differs the most from the ensemble mean (e.g. Giorgi and Mearns 2002). Another way of selecting a subset of models can be based on their degree of independence, a rule of thumb is to consider only one model from each institute (Whetton et al., 2007). Alternatively, Houle et al. (2012) used a cluster analysis in order to classify 86 climate simulations from the CMIP3 archive into 5 subgroups, retaining only a single simulation per subgroup for further analysis. In the case of dynamical downscaling experiments, a reason for

retaining a specific subset of AOGCM from a large ensemble of opportunity may also simply be based on the availability of fields that are necessary for driving an RCM, or some compatibility issues between RCM and AOGCM may also influence the choice of the AOGCMs to be retained in a study.

In addition to the selection of a subset from a large ensemble of simulations, several ways exist for combining simulations from several models for climate-change assessment. For example, a widely used approach is to consider models as equivalent representations of the real climate system, thus using their simulations as equally likely outcomes of the future climate. This can be done by using the arithmetic mean over all models as a measure of the projected climate-change signal; the inter-model spread is then generally interpreted as reflecting the "model uncertainty" (Tebaldi and Knutti 2007) affecting the signal. From a different point of view, some authors argue that since models do not exhibit equal skill at simulating the present climate, they should be weighted based on their performance according to some criteria (Giorgi and Mearns 2002, Tebaldi et al. 2005b, Greene et al. 2006, Räisänen et al. 2010). Such methods allow giving the greatest importance to models that are judged to be more reliable, thus reducing the influence of the less reliable models on the ensemble statistics. The optimal way to combine simulations from a multi-model ensemble is still an open debate (Räisänen, 2007). As a striking example, Christensen et al. (2007) shows that two methods for combining AOGCMs' output into probabilistic climate-change projections (Tebaldi et al. 2005a,b and Greene et al. 2006) lead to results that differ significantly.

In summary, once a large ensemble of opportunity becomes available to the community, the users are exposed to complex choices related to the selection, treatment and combination of these simulations. More precisely, three levels of decisions may be stated as : 1) the pre-selection of simulations of interest to be retained from a large ensemble, 2) the use of downscaling techniques for processing the selected set of simulations, and 3) the mathematical treatment applied for combining the simulations into ensemble statistics or probabilistic projections. In the following, we focus on the uncertainty emerging

from the first level of decision, i.e. the selection of a subset of simulations from a large ensemble.

In Sect. 1.2.1, we briefly describe the multi-model ensemble used in this study and the pre-processing applied to these simulations before further analysis. We also define the ensemble statistics (Sect. 1.2.2), namely the climate-change projections signal and inter-model spread. In Sect. 1.2.3 and 1.2.4, we propose two methods that aim at quantifying the uncertainty related to the selection of the members and the retained models. In Sect. 1.3, the results are illustrated for the case of summer surface air temperature change over North America. We also investigate the effect of the ensemble size, comparing the entire multi-model ensemble and a subset of 11 models, as well as other particular ensembles of smaller size. In Sect. 1.3.3, our analysis leads to a particular representation of the well-known "plume diagram", inspired from (Christensen et al., 2007), that will be seen as "blurred" due to the uncertainty emerging from the selection that affects both the signal and the inter-model spread statistics. Finally, in Sect. 1.3.4, basic constraints that can be applied to a selection process are discussed.

1.2 Experimental Framework

1.2.1 Data and pre-processing

The CMIP3 multi-model dataset has been analysed in the context of the IPCC Fourth Assessment Report (AR4). In the following, we use the simulations performed under the A1B GHGA emission scenario (Tab. 1.1), for the simple reason that it counts the largest number of models and members compared to other scenarios. In the following, the term "multi-model ensemble" (MME) is used to refer to this particular ensemble of 55 simulations. For more information about models' names and specifications, the reader is invited to refer to the PCMDI website at `http://www-pcmdi.llnl.gov`.

The present study focuses on models' results over North America. Each models' historical runs have been combined with the respective projections following the A1B scenario, thus

giving simulations that cover the time period from 1900 to 2100. Climate changes have been calculated over successive 20-year averaging windows, relative to the 1900-1950 average, for each model. Since spatial resolution of the models' atmospheric component varies over a broad range (from 1.1° to 5°), all data were linearly interpolated on the coarsest grid corresponding to that of the GISS-ER model, with a resolution of $(4° \times 5°)$.

1.2.2 Ensemble statistics

The climate-change signal and the inter-model spread are commonly used statistics to summarise the results from a multi-model ensemble of simulations. It is worth noting that the latter statistics is often interpreted as an estimate of the uncertainty of the climate-change signal. To avoid any confusion, we keep the terminology "spread" since "uncertainty" will be used in a different context in the following sections.

Let first ψ_{ik} be any field obtained from one simulation of the i^{th} model in an ensemble of several models. We will consider in the following that the ensemble consists in an array of simulations from several models, with each model being represented by a single realisation. In principle, such an ensemble is not uniquely defined since a number of realisations of each model could have been generated. We refer to any of these possible ensembles by using the k index, which will be discussed in more details in the next two sections. Let us now define the reference past climate (P_{ik}) at time $t = p_o$ as :

$$P_{ik} = \psi_{ik}(x, y, z, t = p_o) \tag{1.1}$$

where x, y and z are the spatial location coordinates. In the present context, p_o corresponds to a time average of the simulation over the reference period from 1900 to 1950. Similarly, the later time climate (F_{ik}) defined over a given 20-year window is written as :

$$F_{ik} = \psi_{ik}(x, y, z, t > p_o) \tag{1.2}$$

where t is larger than p_o, although we focus on the range from 2000 to 2100 in the

following. The climate-change signal (δ_{ik}) for the i^{th} model of the k^{th} ensemble is hence given by

$$\delta_{ik} = F_{ik} - P_{ik}. \tag{1.3}$$

As stated previously, climate-change projections are generally presented as ensemble statistics. In order to obtain the magnitude and the range of the ensemble projections, we define the ensemble-mean climate-change signal ($\overline{\delta_{ik}}^{i}$) and inter-model spread (σ_k) as

$$\overline{\delta_{ik}}^{i} = \overline{F_{ik}}^{i} - \overline{P_{ik}}^{i} \tag{1.4}$$

and

$$\sigma_k = \sqrt{\overline{(\delta_{ik} - \overline{\delta_{ik}}^{i})^2}^{i}} \tag{1.5}$$

respectively, where $\overline{(\cdot)}^{i}$ is the averaging operator over all models in the ensemble.

In the following, we present a general framework that aims at evaluating the uncertainty related to the selection of a sample of simulations from a large ensemble. First, we present in Sect. 1.2.3 the member-sampling approach and evaluate the uncertainty related to the choice of one realisation per model. Then (Sect. 1.2.4), we present the model-sampling approach and evaluate the effect of selecting different subsets of models from the original ensemble.

1.2.3 Member sampling

In this section, we aim at quantifying the uncertainty that is related to the choice of the members when extracting a subset from a large ensemble of simulations. This is done by assuming two constraints to the selection process : 1) one member per model is considered when several are available, and 2) the choice of the models is kept fixed, i.e. is already assumed. We will show how this uncertainty affects the ensemble statistics, namely the ensemble-mean signal and the inter-model spread.

As seen from Tab. 1.1, the 24 CMIP3 models are represented by different and arbitrary

number of members. Let us denote as N_i the number of members available from the i^{th} model. There are hence many ways to form a multi-model array from 24 different models, i.e. $\prod_{i=1}^{24} N_i = 1,360,800$. Each of these variations of the "elected members" that represent the models are associated with an ensemble index, k, which has been introduced in Sect. 1.2.2. Then, for the k^{th} variation of the multi-model array, the ensemble-mean signal and inter-model spread are calculated using (1.4) and (1.5) respectively.

After resampling for a large number (K) of iterations, we perform the following statistics where $\overline{(\cdot)}^k$ is the averaging operator over all generated ensembles. We hence obtain the climate-change signal mean value (Δ_{mem}),

$$\Delta_{mem} = \overline{\overline{\delta_{ik}}^i}^k, \tag{1.6}$$

the uncertainty of the climate-change signal mean value (U^{Δ}_{mem}),

$$U^{\Delta}_{mem} = \sqrt{\overline{(\overline{\delta_{ik}}^i - \overline{\overline{\delta_{ik}}^i}^k)^2}^k}, \tag{1.7}$$

the inter-model spread mean value (Σ_{mem}),

$$\Sigma_{mem} = \overline{\sigma_k}^k \tag{1.8}$$

and the uncertainty of the inter-model spread mean value (U^{Σ}_{mem}),

$$U^{\Sigma}_{mem} = \sqrt{\overline{(\sigma_k - \overline{\sigma_k}^k)^2}^k}. \tag{1.9}$$

In the following, we refer to "member sampling" as the method just described, consisting of randomly choosing one member per model within a multi-model array. It involves a particular assumption on the population from which a subset is drawn. Since we assume a fixed set of models, it should be interpreted as the only opportunity of its kind and hence implicitly as the entire population of the possible modelling approaches. Under

these circumstances, the choice of the members appears as the unique source of sampling uncertainty in the process of selecting one particular subset array.

1.2.4 Model sampling

Since in principle, an infinite number of models could be imagined, let us extend the previous assumption on the nature of the population. We now assume that the multi-model ensemble consists in a representative sample of a larger population. This larger population could be interpreted as including all the possible modelling approaches with a similar level of complexity to the models of the current generation. Based on this assumption, we describe in the following a method for assessing the sampling uncertainty that relates to the choice of the models when constructing a subset from a large ensemble of simulations.

The present method consists of generating many subset arrays by resampling with replacement over the original set of models. Such a method is generally referred to as bootstrap (Wilks 2011). As in Sect. 1.2.3, we constrain the selection process to the use of one member per model that is randomly chosen when several exist; this means that the choice of the members also contributes to the model-sampling uncertainty. This way of generating a multi-model array of size m from a pool of N simulations gives $\binom{N+m-1}{m}$ possibilities. It is worth noting that in combinatorial analysis, this kind of sample is generally called a multiset (Bona, 2006) since one particular element can appear several times. The latter distinction will be discussed in more details in Sect. 1.3.4. In what follows, we however refer to an extracted sample from a larger population as a "subset", which stands for a more general point of view.

By using $m = 24$ models from a pool of $N = 55$ simulations (Tab. 1.1), it is possible to form 7.9×10^{19} different ensembles. However, since the selection of a model is done before the choice of one of its members, an equal probability of occurrence is attributed to 1.6×10^{13} sets differing by at least one model. Naturally, each set may exist under several possible states that differ only by the selected members. This property reflects the fact

that the number of members representing a model does not influence its probability to be drawn from the pool. Moreover, since the resampling is done with model replacement, some of the models may be selected several times while others may not appear at all in a given subset. By assuming the MME to be a representative sample of a larger population, the "model sampling" somehow consists in a generalisation of a classical models' pre-selection phase where each model would not be considered more than once.

For a large number of iterations, one can compute the statistics over the k ensemble index for $\overline{\delta_{ik}}^{i}$ and σ_k, as done in the previous section for the member-sampling approach. We hence obtain similar statistical coefficients for the signal, the spread and their sampling uncertainties labelled as Δ_{mod}, U^{Δ}_{mod}, Σ_{mod} and U^{Σ}_{mod}, corresponding to (1.6) to (1.9) respectively.

1.3 Results

1.3.1 Signal, spread and their uncertainties

We now apply the two approaches described in sections 1.2.3 and 1.2.4 and present the results for the summer surface air temperature over North America. Fig. 1.1 shows the climate-change signal mean value (Δ_{mem} and Δ_{mod}) for three different periods, 2000-2020, 2040-2060 and 2080-2100 (from left to right), relatively to the 1900-1950 climate. The signal is calculated as an average over a large number of ensembles ($K = 2000$) generated using either member (Fig. 1.1a) or model (Fig. 1.1b) sampling approaches. As seen from this figure, the two approaches lead to nearly identical results : an important temperature increase covering the land part of the domain, with a maximum of 5.5°C located on the western part of United States.

The fact that both approaches give very similar results is expected since for the model sampling, all models have the same probability of being chosen. Hence, for a large number of iterations, each model will be chosen an approximately equal number of times, similarly to the member sampling where the 24 models are kept fixed at each iteration.

Since both approaches lead to practically identical results, we use the symbol Δ without subscript in the following to refer to the climate-change signal mean value without regard to the sampling method. It is worth noting that proceeding to a simple arithmetic mean over the entire ensemble of simulations (Tab. 1.1) would have led in effect to a weighted average of the signal, because the relative importance of each model would be determined by their number of available members. It can be shown (Appendix 4.B) that the climate-change mean values presented in Fig. 1.1a and b consist of unweighted ensemble means that give equal relative importance to each of the models in the ensemble, independently of their sample size.

The sampling uncertainty of the signal mean value is displayed in Fig. 1.2 for both approaches (U^{Δ}_{mem} and U^{Δ}_{mod}). For the member sampling (Fig. 1.2a), uncertainty values smaller than 0.06°C cover the domain for the three periods. Patterns display some differences in their shape with time, but the magnitude does not vary substantially. While we interpret this as a measure of the uncertainty related to the choice of members if several are available, this quantity is a manifestation of the natural variability as simulated by the models providing several members to the ensemble.

This measure of uncertainty is expected to underestimate the overall effect of the natural variability that should normally affect the ensemble statistics. In a hypothetical case where the MME would contain a sufficiently large number of members representing each model, our measure of uncertainty would tend asymptotically towards an unbiased estimate of the overall effect of natural variability on the ensemble statistics. This statement is demonstrated in Appendix 1.A through an idealised experiment using synthetic data; this will be also investigated analytically in Chap. 4 (Appendix 4.B). To keep us in perspective, we note that the simulated natural variability can be calculated by using a single but very long climate simulation run under stationary conditions (e.g. Deser et al. 2010), i.e. without external forcing such as GHGA. Under transient boundary conditions such as the present ensemble of simulations, the natural variability is likely to change somewhat over climatic time scales (Sorteberg and Kvamstø, 2006). Hence, the natural

variability may be seen as a time-dependent measure of uncertainty, which can be quantified using the spread between several members generated using a single model with slight differences at the initial conditions. Since both measures are manifestations of the same physical process and that we consider it as a blend-effect from several models, we use the terminology "multi-model natural variability" for describing this specific feature in the following.

Let us now consider Fig. 1.2b where is displayed the uncertainty of the signal mean value that appears from selecting a set of models (U^{Δ}_{mod}). More precisely, this measure of uncertainty consists in the standard deviation of the climate-change ensemble-mean signals (Eq. 1.4) that can be obtained from randomly selecting 24 models with replacement from the MME. The values are considerably larger than those obtained using the member-sampling approach (note the different scales), reaching 0.4°C in the north of Canada. Also, the patterns are consistent in time with an increasing magnitude.

The inter-model spread mean values (Σ_{mem} and Σ_{mod}) are displayed in Fig. 1.3. For the same reason as for Δ, both sampling methods lead to nearly identical results. We will hence adopt the Σ symbol without subscript in the following, for referring to the inter-model spread mean value. One should note the great similarity of the patterns compared to that of U^{Δ}_{mod} (Fig. 1.2b), with a larger magnitude for Σ. Values of Σ reach 1.6°C over the centre of United States and exceed 2°C in the north of Canada. The similarity between Σ and U^{Δ}_{mod} can easily be understood since, by analogy to the standard error relationship (von Storch and Zwiers, 1999), the model-sampling uncertainty that affects the signal should be proportional to the inter-model spread and inversely proportional to the square root of the number of models.

Finally, the uncertainties of the inter-model spread mean value (U^{Σ}_{mem} and U^{Σ}_{mod}) are shown in Fig. 1.4. For the member sampling (U^{Σ}_{mem}), patterns vary with time, without any general changes in magnitude, as was the case for the uncertainty of the signal due to the selection of the members (Fig. 1.2a). Relatively small values of U^{Σ}_{mem} cover the continental region, in a range between 0.05°C and 0.07°C. Based on the results of

Appendix 1.A stating that the member-sampling uncertainty of the signal (Fig. 1.2a) underestimates the real effect of the multi-model natural variability, the member-sampling uncertainty that affects the inter-model spread is also expected to be underestimated due to the relatively small number of members available for each model in the ensemble. The model-sampling uncertainty of the inter-model spread (U^{Σ}_{mod}) is, as expected, larger than U^{Σ}_{mem}, with a maximum of 0.6°C in the north of Canada. One should note some similarity between the model-sampling uncertainty of the inter-model spread (Fig. 1.4b) and the inter-model spread mean value (Fig. 1.3). These two quantities are related in a similar way as are the uncertainty of the signal mean value and the inter-model spread mean value through the standard error relationship. For instance, the expected error of the variance estimator is proportional to the square of the population variance for a dataset consisting in independent and identically distributed normal random variables (von Storch and Zwiers, 1999).

1.3.2 Smaller ensemble of opportunity

The previous section presented results using all of the 24 models that are available in the ensemble. One issue in assessing the member-sampling uncertainty is that 13 of the 24 models are represented by only one member. For each subset array obtained with the member-sampling approach, these 13 models do not allow any possibility of varying the elected members, thus increasing artificially the apparent stability of the ensemble statistics and hence decreasing the magnitude of the resulting perceived uncertainty. An approach that aims at minimising this issue is simply to remove from the ensemble the 13 models represented by only one member. By the same logic, one could also choose to retain only the models with the largest number of members after removing those with a sample size that is smaller than some predefined threshold. It is worth noting that this specific pattern of models' pre-selection to form smaller subsets allows maximising the number of possible ensembles that can be formed by using the member-sampling approach. A side effect is that it compromises the diversity of the models in the calculation of the member-sampling uncertainty. According to Appendix 1.A, the

11-model subset that is based on the models with at least two realisations consists in an educated guess in order to minimise the side effects from an unequal and small number of members across the models. It is hence expected that a small systematic bias will remain between our measure of the member-sampling uncertainty and the expected value of multi-model natural variability. In the following, we present results in a similar way as in Sect. 1.3.1, but for the 11-model ensemble where each model is represented by at least two members.

Fig. 1.5a shows the signal mean value for the surface air temperature obtained using the 11-model ensemble. Compared to the signal mean value calculated from the entire 24-model ensemble (Fig. 1.1), the signal extracted from the 11-model ensemble has a maximum that is located nearer to the west coast of United States, with slightly weaker intensity.

For the member-sampling uncertainty of the signal mean value (Fig. 1.5b), the patterns are very similar to that of Fig. 1.2a, but with more than twice its intensity; since the models with only one member have been removed from the ensemble, the member-sampling uncertainty of the signal mean value originates from same variations of the elected members as for the entire 24-model ensemble. The member-sampling uncertainty obtained for the 11-model ensemble consists in a multi-model blend of natural variability, but its overall effect on the signal mean value is expected to be slightly underestimated.

For the model-sampling uncertainty of the signal mean value, Fig. 1.5c shows an increase of the uncertainty compared to Fig. 1.2b. The values reach approximately 0.15°C over central United States. This increase of the uncertainty is mainly due to the larger standard error of the mean when using a smaller sample of models. A second contribution is that shown in Fig. 1.5b, since the member-sampling uncertainty is implicitly included in that of the model sampling.

In order to analyse further the impact of the sampling uncertainty on the ensemble statistics, we introduce here a measure of relative uncertainty, calculated as U^{Δ}_{mem}/Δ

and U^{Δ}_{mod}/Δ, as shown in Fig. 1.6a and b for member and model sampling, respectively. It can be seen from Fig. 1.6a that the member-sampling relative uncertainty decreases with time due to the fact that the intensity of the member-sampling uncertainty is approximately constant with time (Fig. 1.5b) while the climate-change signal mean value increases. For the 2000-2020 period (Fig. 1.6a, left panel), the values are generally smaller than 10% over the continental region, while two strong maxima ($> 15\%$) are located in the north of Canada and near Greenland. The model-sampling relative uncertainty of the signal for the 2000-2020 period (Fig. 1.6b, left panel) gives values that can reach 20% over the continental regions, while maxima over the Pacific Ocean and near Greenland are larger than 35%. With time, the model-sampling relative uncertainty of the signal decreases briefly after the 2000-2020 period, but remains approximately constant from 2020-2040 to 2080-2100.

Let us now take a look to the inter-model spread mean value and its components of uncertainty for the 11-model ensemble (Fig. 1.7). As for the signal mean value, the patterns of the inter-model spread (Fig. 1.7a) are substantially different from those obtained using the 24-model ensemble (Fig. 1.3). The magnitude of the inter-model spread for the 11-model ensemble is smaller mostly over continental regions, while over ocean it tends to exhibit similar values. For the member-sampling uncertainty of the spread mean value (Fig. 1.7b), patterns vary with time, with some persisting features such as the maximum over United States. The model-sampling uncertainty of the inter-model spread (Fig. 1.7c) grows with time as for the spread mean values for both ensembles of models (Fig. 1.7a and Fig. 1.4b).

In Fig. 1.8 is shown the relative uncertainty of the spread mean value, written as U^{Σ}_{mem}/Σ and U^{Σ}_{mod}/Σ for the member and model sampling respectively. For the member sampling, the relative uncertainty decreases with time since the absolute uncertainty (Fig. 1.7b) does not change so much in intensity while the spread mean value (Fig. 1.7a) considerably increases with time. Finally, the relative uncertainty due to the sampling of the models (Fig. 1.8b) is nearly constant with time, due to the fact that the uncertainty component

increases at a similar rate as the spread mean value.

To summarise our results, we present in the following the different components of uncertainty as function of the size of the ensemble subset, for one grid point centred over the Québec province in Canada. We use particular ensembles of different sizes chosen by removing models represented by less than 2, 3, 4 and 5 members, resulting in ensembles formed by 11, 10, 5 and 3 models, respectively. This particular pattern for selecting the models allows optimising the sampling of the natural variability in smaller multi-model ensembles. Such an approach tends to maximise the measure of the member-sampling uncertainty by minimising its systematic bias (due to an insufficient number of members). However, we note that it does not maximise the intensity of the member-sampling uncertainty from the physical point of view, given that important inter-model differences exist in the simulated natural variability. The different components of uncertainty are shown in Fig. 1.9 for the three time periods. Reducing the size of the subset ensemble increases the member-sampling uncertainty, similarly for both the signal and the spread. While the member-sampling uncertainty exhibit a similar magnitude for both signal and spread, at all times, the model-sampling uncertainty increases with time at a similar rate for the signal than for the spread.

Finally, the components of relative uncertainty as function of the ensemble size are shown in Fig. 1.10. The relative uncertainty increases when reducing the ensemble size, at a faster rate for the uncertainty of the spread compared to that of the signal, for both member and model sampling uncertainties. The member-sampling relative uncertainty of signal and spread diminishes with time, while the model-sampling relative uncertainty shows some decreases from 2000-2020 to 2040-2060 and remains approximately constant until 2080-2100. This stability with time of the relative uncertainty is due to the fact U^{Δ}_{mod} and U^{Σ}_{mod} increase at a rate that is partly balanced by Δ and Σ respectively. For U^{Σ}_{mod}/Σ, this balance is somewhat expected since U^{Σ}_{mod} depends on Σ. For U^{Δ}_{mod}/Δ however, the balance occurs rather by chance since U^{Δ}_{mod} depends on Σ while Δ depends on the intensity of the GHGA forcing.

1.3.3 Revisiting the plume diagram

In this section, we extend the concept of the sampling uncertainties to the construction of plume diagrams presented in Christensen et al. (2007). From Sect. 1.2.4, the model-sampling approach is applied to calculate the signal and inter-model spread mean values (Δ and Σ) with their respective uncertainties (U^{Δ}_{mod} and U^{Σ}_{mod}) over the entire 200-year period; we also calculate U^{Δ}_{mem} and U^{Σ}_{mem} as described in Sect. 1.2.3.

Let us now present the plume diagrams for the surface air temperature field over a grid point centred over the Québec province of Canada, using the 24-model ensemble (Fig. 1.11a), the 11-model (Fig. 1.11b) and the 5-model ensemble (Fig. 1.11c). The signal mean value (Δ) is displayed as the blue full line; the signal uncertainty that is due to the sampling of the models (U^{Δ}_{mod}) is drawn as blue dashed lines calculated as $\Delta \pm 2 \times U^{\Delta}_{mod}$, and similarly for the sampling of the members as blue dotted lines ($\Delta \pm 2 \times U^{\Delta}_{mem}$). The upper and lower boundaries of the ensemble envelope (red full lines) are calculated by using the signal and inter-model spread mean values combined as $\Delta \pm 2 \times \Sigma$. The model-sampling uncertainty affecting the envelope boundaries is given by the four red dashed lines and calculated as $\Delta \pm 2 \times \Sigma \pm 2 \times U^{\Sigma}_{mod}$; similarly, the member-sampling uncertainty is displayed as red dotted lines using $\Delta \pm 2 \times \Sigma \pm 2 \times U^{\Sigma}_{mem}$.

The plume diagram displayed in Fig. 1.11a is obtained using the entire 24-model ensemble and shows a surface air temperature increase of 4°C in the signal mean at the horizon 2080-2100. The signals obtained from different ensemble sizes differ slightly, with values of 3.5°C and 3.3°C for the 11-model and 5-model subsets, respectively. It should be noted that it is not an effect due to the ensemble size, but rather a consequence of the particular choices of the models forming the ensembles.

For the same time horizon of 2080-2100, the signal has a model-sampling uncertainty of 0.5, 0.6 and 0.8°C for the 24-, 11- and 5-model ensembles, respectively. This component of uncertainty describes the stability of the signal coefficient for arbitrary ensembles of the prescribed sizes, under the assumption that the MME consists in a representative sample

of a larger population that would consider all the possible modelling approaches with a similar level of complexity than those that are currently available in the pool. Indeed, if new models were available in the pool for example by including more processes and having finer resolution, there are reasons to believe that the inter-model spread among models would likely increase (Knutti, 2010), hence inducing a corresponding increase of the perceived uncertainties.

Similarly to the signal, the inter-model spread has different values depending on which models are chosen to form the ensemble, with values of 2.3, 2.0 and 1.8°C at the horizon 2080-2100 for the three ensemble with 24, 11 and 5 models. As discussed above, the uncertainty of the signal due to the sampling of models increases when reducing the ensemble size, but also is proportional to Σ as it is generally the case for standard errors. The uncertainty of the inter-model spread is 0.2, 0.2 and 0.4°C for the 24-, 11- and 5-model ensembles, respectively. In both cases of signal and inter-model spread, the uncertainty due to the sampling of the members has very little impact on the statistics (blue and red dotted lines, respectively) for the 24-model ensemble, but it increases significantly when reducing the ensemble size. As noted previously, the latter source of uncertainty is probably underestimated, particularly for the 24-model ensemble, due to the very poor sampling of the members compared to the number of models within the ensemble.

The plume diagrams shown in Fig. 1.11 characterise one variable over a particular location. The general idea to be retained from these results however is that these plume diagrams appear as "blurred" in both their mean and spread components. This blurring is the perceived error in the ensemble statistics and aims at representing the uncertainty that face experts when selecting a subset from a large ensemble of simulations. It consists in a manifestation of known sources of uncertainty, in the present case the natural variability and the inter-model spread.

1.3.4 Constraining the selection process

Like any decision process that is not based on pure randomness, choosing a set of simulations from a large ensemble should involve well-defined constraints. While such constraints may be very specific to the matter of a study, others are of common use in climate sciences. One rather popular constraint is the use of one realisation per model if several are available (e.g. Bombardi and Carvalho 2011, Peings and Douville 2010, Räisänen et al. 2010), which allows to considerably reduce the size of an ensemble while retaining much of the information relative to the model uncertainty (inter-model spread). One may also think of more complex constraints, for example based on model performance (Gleckler et al., 2008), institutional independence (Whetton et al., 2007) or clusters in the phase space (Houle et al., 2012).

Other constraints may, at first sight, appear as implicit but can be relaxed for more generality. For example, the model-sampling technique that has been applied throughout this article employs model replacement. However, in a real expert-based process of selecting a multi-model array of simulations, a same model is generally not included more than once. Allowing for model replacement in the resampling technique has been intimately related to the assumption that the MME is a representative sample of a larger population. On the other hand, if one assumes that the MME is not representative of any larger but rather consists in the entire population of modelling approaches, the selection of a set of models should be done without replacement. In what follows, we investigate the effect of choosing one of these two assumptions about the nature of an ensemble of opportunity through an experiment based on synthetic data.

Let us consider an artificial 24-model array where each model is represented by a single number. By simplicity, we assume a unique member to be available for each of the models. This array is generated using a random number generator based on a normal distribution with zero mean and unit variance. We hence apply the model-sampling approach as described in Sect. 1.2.4 for $K = 2000$ iterations and repeat this procedure for each

sample size m from 24 to 2. For each drawn subset of a given size, the ensemble mean is first calculated and next the standard deviation over all the means (previously denoted as U^{Δ}_{mod}, according to Eq. 1.7) is calculated. In a similar way, we repeat the previously described technique over the same initial 24-model array but with the difference that we use a "no-replacement constraint" throughout the resampling procedure, unlike what is normally done for bootstrapping techniques.

The results are shown in Fig. 1.12a where the blue and green curves represent the perceived error of the mean depending on the version of the model-sampling technique, i.e. with and without replacement, respectively. Both curves are normalised using the standard deviation of the initial sample, which in practice differs from 1 because the 24-model ensemble is of finite size. For reference, a red curve has been added corresponding to the well-known standard error law in its normalised form, i.e. $1/\sqrt{m}$. As seen from the figure, the uncertainty of model sampling with replacement agrees fairly well with the standard error relationship. A subtle underestimation however appears in intensity, which is due to the finite size of the initial sample (i.e. $m = 24$).

According to Fig. 1.12a, the underlying assumption about the nature of the ensemble of opportunity plays an important role in interpreting the ensemble statistics from the point of view of their sampling uncertainty. The important differences between the blue and green curves can be related to the number of combinations that can be formed from the initial pool. While $\binom{N+m-1}{m}$ multisets (with replacement) of size m can be formed from a pool of N elements, $\binom{N}{m}$ subsets (without replacement) are possible. For convenience, Fig. 1.13a shows the number of combinations that are possible for multisets (blue) and subsets (green). The trivial case consists in $m = 24$ where the extracted sample has the same size as the entire pool. Thus, it leaves a single possibility of forming an ensemble by considering all of the models. For $m = 24$ in Fig. 1.12a, a null value of uncertainty is hence attributed to the selection of a group of models since the extracted sample consists in the entire population. For an equal sample size, model sampling with replacement allows for 1.6×10^{13} different combinations of models. In the latter number of combinations,

all but one involve at least two replicates of a same model and these combinations result in a value of 0.2 for the uncertainty of the mean.

It is worth noting that the possible ensembles that can be formed without replacement are included in those for the case that allows for model replacement. The blue curve in Fig. 1.13a hence represents the total number of ensembles that can be formed both with and without replacement, while the green curve accounts only for the latter. By analysing these two curves, it can be seen that the relative importance of the number of possible ensembles formed without replacement increases relatively to the total number of ensembles. This can be seen more clearly in Fig. 1.13b where is shown the ratio between the number of subsets and multisets. This ratio diminishes rapidly when increasing the sample size, the number of subsets (without replacement) representing 43% of the total number of combinations at $m = 5$, while it shows 2% for $m = 10$. Since, when decreasing the sample size, the combinations formed without replacement represent an increasing proportion of the total number of combinations, this necessarily leads to converging errors of the mean as seen in Fig. 1.12a. The fact that both measures of uncertainty converge when decreasing the sample size can be seen in Fig. 1.12b where is shown the ratio between the error emerging from resampling without and with replacement. The error ratio decreases monotonically with increasing the sample size, resampling without replacement representing approximately 90% of the error obtained by allowing replacement for $m = 5$, while it represents 40% for $m = 20$.

In summary, introducing constraints to the selection process along with a specific assumption about the nature of the population are considerations that play an important role in characterising the uncertainty related to experts' decisions. Moreover, we note that a systematic application of a constraint may reveal some patterns in the decision process. For example, in Sect. 1.3.2, we constrained our selection of smaller subsets by advocating models with largest sample sizes in order to minimise the systematic biases affecting the member-sampling uncertainty. A worrying pattern that could emerge from such a strategy could consists in a "liberal picture" of the multi-model natural variabi-

lity, which tends to over-represent the wealthiest research centres with larger resources allowing performing a larger number of simulations. Another pattern of selection that will be investigated in Chap. 2 consists in constraining the selection to the models that have been developed by different modelling centres. Such an approach reveals important benefits by limiting the occurrence of uninformative consensus that contaminate the message conveyed by an ensemble of opportunity.

1.4 Discussion and conclusions

Climate-change projections are mainly based on AOGCMs' simulations that are forced by increasing concentration of GHGA in the atmosphere over periods extending from decades to centuries. It is well known that ensembles of such simulations generally lead to a broad range of climate-change projection when using several models, which may differ from a structural point of view, additionally to differences in the tuning of weakly constrained parameters and in the numerical approximations used in the discretisation of the equations. Several internationally coordinated projects have been set up over the last decade in order to compare models' results and quantify the inter-model spread that is often interpreted as the uncertainty in modelling the climate system. In general, such ensembles do sample the differences in modelling approaches, but neither in a random nor in a systematic manner. Often called ensembles of opportunity, these are formed in a rather open way : the resulting sample of simulations highly depends on the fact that research centres are free to participate by delivering an arbitrary number of members generated using one or several models or versions.

Ensembles of opportunity are important for informing the public, the scientific community and the policy makers about future climate changes. Additionally to their direct use in climate-change assessments, these simulations are often used as an input to other kinds of models, such as Regional Climate Models (RCMs; Rummukainen 2010) and Statistical Downscaling Models (e.g. Dibike et al. 2008). However, many centres that use the output data from an ensemble of opportunity are constrained to process only

a small set of simulations compared to the available ensemble. For example, the dynamical downscaling of AOGCM simulations using an RCM involves large computational resources due to the use of high resolutions and long time periods, additionally to the treatment needed for preparing the driving fields. The selection of a subset of simulations from a large ensemble is generally based on experts' judgement, depending upon the goals of the study. A rather popular choice for reducing the size of a large ensemble is to use only one member per model when several are available. This choice is generally supported by the assumption that the inter-model spread is more important than the simulated natural variability when using simulations over long time periods (e.g. several decades). Other types of decisions are more specific to the selection of the models to be part of the new ensemble, for example based on their simulation of specific climate features, model performance, institutional independence, compatibility issues or simply based on the availability of particular fields of interest (e.g. needed for driving an RCM). As stated in introduction, experts generally face up three levels of decision when using the data from a large ensemble of opportunity : the pre-selection of a set of simulations from the available ensemble, the use of other types of models for processing the AOGCMs' output and the combination of the simulations into ensemble statistics or probabilistic projections. In this chapter, we aimed at quantifying the uncertainty related to the first level of experts' decisions. The second and third levels are simplified by not using any other kinds of models (e.g. RCMs) and by calculating common ensemble statistics, namely the ensemble mean climate-change signal and the inter-model spread.

The process of selecting a subset of simulations from a large ensemble has been investigated by considering the selection of both the members and the models. We first defined the member-sampling approach that is based on two basic constraints : one member per model is retained when several are available and the selection of the models forming the ensemble is kept fixed. It results in a large number of possible multi-model arrays that differ only by the selection of members. From a more general point of view, the member-sampling approach assumes the pre-selected sample of models as the one and only opportunity of its kind and hence it consists in the entire population of modelling

approaches. No sampling uncertainty is hence attributed to the selection of the models while the only source resides in the choice of the member (if several are available) that represents each model. By resampling over a large number of multi-model arrays, we obtained the member-sampling uncertainty of the ensemble statistics, which is a manifestation of the models' natural variability.

On the other hand, the model-sampling approach is also constrained by the selection of one member per model if several are available, but without the second constraint, that is the selection of the models is not fixed *a priori*. By resampling over several multi-model arrays, this approach allows to assess the sampling uncertainty that is due to both the choice of the models and the members. Additionally to the fact that the model sampling accounts for the choice of the models, it is more general than the member-sampling approach in its underlying assumption about the nature of the ensemble of opportunity. More precisely, it assumes that the MME is not unique but rather consists in a representative sample taken from a larger population of possible modelling approaches. In the present experiment, this representativeness emerges from allowing model replacement in the sampling method. It results in a model-sampling uncertainty that is a direct manifestation of the inter-model spread through the well-known standard error relationship.

As seen from the results, the member-sampling uncertainty shows very small values for both signal and spread, especially when using all of the 24 models. We emphasise on the fact that the member-sampling uncertainty underestimates the real natural variability in the ensemble statistics, as shown in an alternative experiment using synthetic data (Appendix 1.A). This underestimation is mainly due to the small and unequal number of members representing each model in the ensemble. We aimed at reducing this systematic bias by removing from the ensemble the models with less than 2, 3, 4, and 5 members, thus increasing the minimal number of members that are available for the considered models. One drawback from not considering some of the models consists in a reduction of models' diversity in the sampled natural variability. As expected for both the signal

and the spread, the member-sampling uncertainty increases as function of the decreasing number of models in the ensemble. The uncertainty also remains approximately constant with time in absolute terms, but decreases when analysed as relatively to the ensemble mean signal and inter-model spread, which both increase with time.

In the results obtained from the use of the model-sampling approach, the uncertainty for both signal and spread has been shown to increase when diminishing the number of models in the sample, as it is generally the case with standard errors. It also appeared to increase with time, since the model-sampling uncertainty necessarily depends on the inter-model spread. The model-sampling relative uncertainty displayed some tendency to remain constant with time, which seems to occur somehow by chance for the signal since the signal and the inter-model spread are not directly related : the strength of the signal depends on the magnitude of the GHGA emissions while the inter-model spread depends mainly on the structural differences between models and their different response to changes in forcing. For the relative uncertainty of the inter-model spread, this balance can be expected since the model-sampling uncertainty of the spread depends on the spread itself.

The member and model sampling uncertainties have been used in the construction of a plume diagram, where the signal mean and inter-model spread appear as "blurred" features. The thickness of the model-sampling uncertainty envelopes affecting the signal and the spread necessarily depends on both the inter-model spread and the number of models involved in the ensemble, additionally to a contribution from the member sampling. Taken alone, the member-sampling envelope of uncertainty depends on the number of models forming the ensemble and consists in a blend of the natural variability as simulated by several different models. This blend is attenuated compared to the real effect of the natural variability that is expected from the use of an ensemble where each model would be represented by a sufficiently large number of members. As will be seen in Chap. 4, the extent to which the member-sampling uncertainty underestimates the real value of the multi-model natural variability mainly depends on the minimal number

of members that represent a model in the ensemble.

Our experimental framework aimed at quantifying the uncertainty range resulting from the choices made by data users when selecting a set of simulations from a large ensemble. It is important to note that this kind of uncertainty should not be seen as a supplementary source to those that are currently known such as GHGA emission pathways, inter-model differences and natural variability (see Foley 2010 for a review). Rather, the sampling uncertainties should be interpreted as different manifestations of these known sources. Moreover, our perception of these sources can be altered through the selection process depending on the constraints that are involved and the assumption about the nature of the ensemble of opportunity. For instance, the member-sampling uncertainty underestimates the multi-model natural variability while the model-sampling uncertainty is a direct consequence of the inter-model spread (model uncertainty) through the standard error relationship. It is worth noting that a model selection based on the assumption that the MME consists in the entire population of modelling approaches would not involve any model replacement and hence the perceived sampling uncertainty would consist in an underestimation of what is expected from the standard error relationship. Of course, the present approach for assessing the sampling uncertainties could be extended to more complex ensemble statistics than means and variances, for example by using quantiles. We also note that the previous results are conditional to one specific emission scenario and hence that the sampling uncertainties could be extended to the scenario dimension as well, additively to the models and the members.

A variety of possible choices that are left to the user when selecting simulations from a large ensemble has been explored in this chapter by using resampling methods based on a set of prior constraints. The main constraint that consists in retaining a single member per model reflects a typical decision made by experts in order to efficiently reduce the size of a large ensemble of opportunity. While the choice of the members is often done randomly in real-life applications, we acknowledge that this method is a simplified representation of a real expert-based process of selecting a set of models from a large

ensemble. From that point of view, our experimental framework could be extended for better representing such a process. For example, one could add constraints to the selection process, such as by forbidding several models developed by a same research centre to be part of a same ensemble. Our approach could be also used to seek for ensembles with special characteristics, for example by maximising the inter-model spread for a given ensemble size. Though some similarities could be noted between the latter application and clustering methods (Houle et al., 2012), considering both approaches in complementarity could provide a powerful framework for investigating the simulations' pre-selection problem. On the other hand, however, it is always worth questioning the potentially diminishing return of implementing complex and expensive strategies for selecting a set of simulations. This is especially true when, as seen previously, the sampling uncertainty related to the choice of an ensemble is smaller than what should be expected from the known sources.

In the application of the model-sampling technique, the generation of a particular ensemble could be related to the use of an arbitrary weighting procedure, in the sense that some of the models may be accounted for several times while others are not considered at all in the calculation of the ensemble mean. By applying the model-sampling technique with several iterations, it however appears as contradictory with a potential assessment of the uncertainty related to such weighting procedures. The important difference is that the model-sampling technique weights the models in an unconstrained manner, i.e. randomly and uniformly, while by definition, the common weighting procedures attribute weights according to specific physical constraints (e.g. Allen and Ingram 2002). While a robust constraint should not involve a large dependence of the weighted results onto the ensemble under consideration, applying the model-sampling technique according to a given physical constraint could allow to quantify the sampling uncertainty related to that specific weighting procedure, thus providing a comprehensive measure of robustness.

Another possible application involves the member-sampling approach that could be used for investigating the overall effect of the natural variability that arises from a combina-

tion of several models. As said previously, the uncertainty due to the selection of one member per model results in an underestimation of the real effect of the simulated natural variability due to the insufficient and unequal number of members representing each model. As shown in Appendix 1.A, such a systematic bias can be reduced by applying a correction factor based on synthetic experiments results. On the other hand, the analysis of variance (ANOVA) is a popular approach used for decomposing the variability of a system into several components, such as the model uncertainty and the natural variability. However, such an approach is not always suitable for the case of unequal sample sizes, especially when the smallest samples count very few elements.

From a more general point of view, both the selection of a set of simulations and the application of weighting procedures can be understood as expert-based samplings that are applied to an already existing ensemble. We note that this supplementary sampling should be well distinguished from the initial sampling of the ensemble. Often described as neither random or systematic, the initial sampling of an ensemble of opportunity can be interpreted as a "natural pre-selection" between several modelling approaches, a process by which a better representation is given to modelling centres that can afford the delivering of a larger number of simulations. On the other hand, an expert-based selection is often applied due to limited resources for handling all of the available data. In such a case, the reduction of a large ensemble has to be carefully done by minimising any potential loss of information that could serve the purpose of the study. Another reason for applying an expert-based selection is the aim at "correcting" some uneven characteristics that appear in the initial sampling of an ensemble, such as by filtering out supplementary models developed by a same institute, a method that might help to reduce the occurrence of uninformative consensus between the models' results.

The two main assumptions about the nature of the population related to an ensemble of opportunity that have been investigated through this work are at the basis of a majority of studies in climate sciences, as well as in other fields. While these points of view can be argued for, they remain very specific to the ensemble at hand by either considering

it as the entire population or by projecting it toward a larger ensemble but with similar distributional characteristics. In consequence, it hides an important part of the problem by not questioning the "neither random or systematic" nature of the sampling process by which an ensemble of opportunity has been formed. In order to provide a clearer picture of this important issue, one should think of a third assumption about the nature of an ensemble of opportunity by which it is not representative and likely to be biased from an idealised population of modelling approaches. While the task of defining such a population may seem out of reach, some realistic considerations can be made at least in theory; it should be formed by mostly independent climate models in order to allow some cancellation of the models' respective biases with an increasing sample size. Such a task should necessarily be undertaken by filtering the multi-model ensemble according to robust constraints of selection or weighting procedures, which are far from making broad consensus within the climate-science community.

Appendix 1.A : Perfect-ensemble experiment for bias correction in the statistics related to an unbalanced design

In this appendix, we present an idealised experiment conducted with synthetic data in order to evaluate how the uncertainty emerging from the member-sampling approach is affected by the "imperfect geometry" of a multi-model ensemble, due to the fact that there is a varying number of members for different models.

Let us consider P as an ensemble of simulations composed of 24 models, where each model would be represented with 24 realisations. The simulations can be arranged as a matrix where the models are distributed horizontally and the realisations vertically. We refer to this matrix as a "perfect ensemble", since every model would have an equal number of realisations and because the number of members for each model is sufficiently large to offer a relatively good sample of each model's natural variability. Structurally, the multi-model ensemble (MME) shown in Tab. 1.1 can be considered a subset of P, as shown in Fig. 1.14, where the black elements represent the available simulations from the MME and the white elements the simulations that are missing compared to P. The subset is denoted as I and is referred to as the "imperfect ensemble".

A perfect-ensemble experiment will be realised based on synthetic data and hence does not imply any real data from climate models. From the concept of the perfect (P) and imperfect (I) ensembles, as previously defined, random processes are used to emulate the climate models' database. The methodology can be summarised through the following steps :

1. Construct the P matrix where each element P_{ij} is generated using the following statistical model :

$$P_{ij} = a_i + b_{ij} \tag{1.10}$$

where the two components on the right-hand side consist of normally distribu-

ted random processes characterised as :

$$a_i \sim N(\mu = 0, \sigma^2 = 100) \tag{1.11}$$

$$b_{ij} \sim N(\mu = 0, \sigma^2 = 25) \tag{1.12}$$

where μ and σ^2 are the mean and the variance defining each of the two processes. Thus, every element P_{ij} mimics a climate-change signal simulated by the j^{th} member of the i^{th} model. The values of variance in (1.11) and (1.12) have been chosen arbitrarily but their relative magnitudes aim at emulating the inter-model spread ($\sigma = 10$) and the natural variability of individual models ($\sigma = 5$). We assumed $I = 24$ models, and for each model, $J = 24$ members are generated.

2. Construct I by applying the MME mask (Fig. 1.14) to P.
3. Apply the member-sampling approach to both sets (P and I), with 2000 iterations, to obtain the member-sampling uncertainty. The sampling uncertainty of the mean is calculated from P and I and denoted with $U_{mem}^{\Delta,P}$ and $U_{mem}^{\Delta,I}$ respectively.
4. Repeat steps 1 to 3 several times (1000 iterations), where each iteration consists in a new initialisation of the P matrix according to the model described by (1.10) to (1.12).
5. Repeat step 4 using different definitions of the P and I ensembles. Similarly to the selection pattern that is applied in this chapter, we successively remove the models with the smallest number of members, i.e. from the right to left in Fig. 1.14. The selection pattern is applied to P and I (and similarly to the mask), leading to ensembles of 11, 10, 5 and 3 models, thus maximising the number of available elements for each reduction along the model axis.

We now consider the results obtained for the 11-model ensemble. In Fig. 1.15 are shown

the distributions for the sampling uncertainty of the ensemble mean, with normalised frequencies, for the perfect ($U_{mem}^{\Delta,P}$, left panel) and imperfect ($U_{mem}^{\Delta,I}$, right panel) cases. These distributions represent the range of values taken by the uncertainty across the 1000 experiments, each one being characterised by a new initialisation of the P matrix. For the perfect case, the distribution of the uncertainty is not centred on the red line that indicates the expected value of uncertainty. This expected value consists in the standard error of the mean that emerges from natural variability, hence $\sigma/\sqrt{m} = 1.5$, where $\sigma = 5$ from (1.12) and $m = 11$. The distribution is slightly biased toward smaller values, meaning that 24 members are not enough for the matrix to be strictly perfect. Due to the large computational cost of the present experiment, we still assume that P is a perfect matrix by neglecting this small bias.

It can be seen in Fig. 1.15 that the distribution of the member-sampling uncertainty in the 11-model ensemble mean for the imperfect case (right panel) is characterised by changes in its parameters, namely the location (mean) and the scale (standard deviation). When compared to the perfect case (left panel), the imperfect ensemble shows a systematic bias toward smaller values and its scale is larger.

Let us now introduce a bias-correction factor (G) that characterises the transformation of the distribution from the imperfect to the perfect case. This correction factor can be written as follows :

$$G = \frac{U_{mem}^{\Delta,P}}{U_{mem}^{\Delta,I}} \tag{1.13}$$

consisting in the ratio between the member-sampling uncertainty obtained from the perfect and the imperfect matrix, for each of the 1000 iterations (step 4) and ensemble sizes (step 5). The distribution of G as function of the number of models forming the ensemble is shown in Fig. 1.16. When reducing the number of models and thus increasing the number of available elements relatively to the perfect matrix, the distribution of the bias-correction factor moves to the left and tends to be centred over 1. For the 24-model ensemble, G is distributed around 1.7 and stands far away from the other ensemble sizes, which are centred slightly higher than 1. While the distributions for the

11- and 10-model ensembles are quite similar, the 5 and 3-model ensembles show some displacement toward unity. However, despite that the correction factor gets closer to 1, reducing the dimension of the model axis leads to some loses in the model diversity of the ensemble in its sampling of the natural variability. The 11-model ensemble seems to be the best compromise between having a bias-correction factor near of 1 and by keeping the largest amount of information about the model diversity in natural variability. It is worth noting that the results for G do not vary when changing the input parameters (μ and σ^2) of the model described by (1.10) to (1.12).

From a practical point of view, one could use the G factor in order to de-skew (by simple multiplication) the uncertainty emerging from the member-sampling approach obtained in the present chapter. Considering the present perfect-ensemble experiment, the MME shown in Tab. 1.1 can be seen as one realisation of the I matrix over 1000 (step 4). It results that the correct value of G for the particular case of CMIP3 is unknown, but can be expected to be part of the distributions shown in Fig. 1.16. An educated guess for the choice of G would be to use the value with the highest frequency of occurrence in the distribution, for example, $G = 1.7$ and 1.14 for the 24- and 11-model ensembles, respectively.

Tab. 1.1: Multi-model ensemble formed by 24 AOGCMs taken from the PCMDI archive, which provide climate-change projections based on the A1B emission scenario. The sample size (N_i) corresponds to the number of members available for the i^{th} model for a total of 55 runs. For more information about models' names and specifications, the reader is invited to refer to the PCMDI website at `http://www-pcmdi.llnl.gov`.

i	**Models**	**Sample size** (N_i)
1	BCCR-BCM2.0	1
2	CCSM3	7
3	CGCM3.1(T47)	5
4	CGCM3.1(T63)	1
5	CNRM-CM3	1
6	CSIRO-MK3.0	1
7	CSIRO-MK3.5	1
8	ECHAM5/MPI-OM	4
9	ECHO-G	3
10	FGOALS-g1.0	3
11	GFDL-CM2.0	1
12	GFDL-CM2.1	1
13	GISS-AOM	2
14	GISS-EH	3
15	GISS-ER	4
16	INGV-ECHAM4	1
17	INM-CM3.0	1
18	IPSL-CM4	1
19	MIROC3.2(hires)	1
20	MIROC3.2(medres)	3
21	MRI-CGCM2.3.2	5
22	PCM	3
23	UKMO-HadCM3	1
24	UKMO-HadGEM1	1

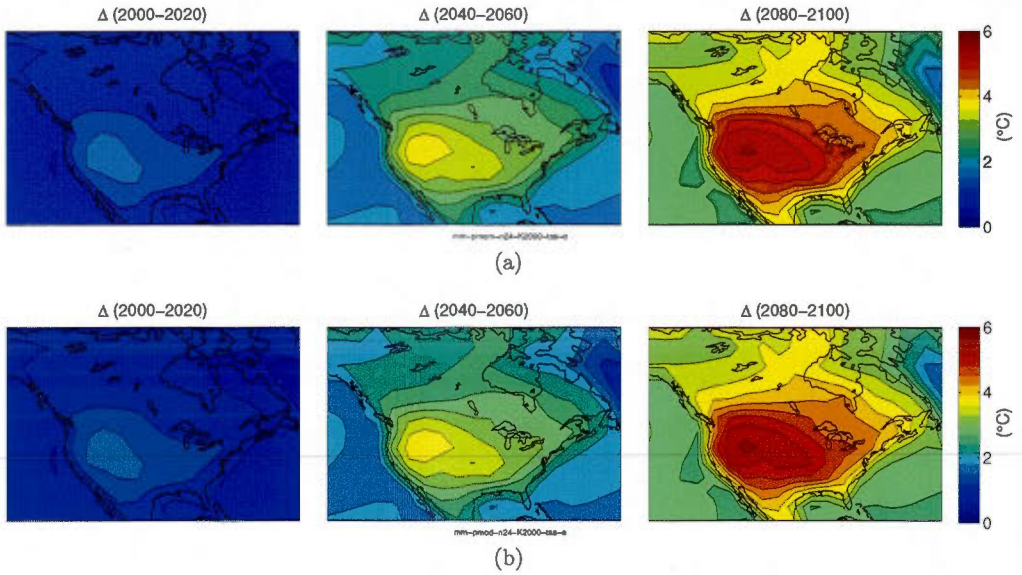


Fig. 1.1: Signal mean value of climate change calculated using a) the member sampling (Δ_{mem}) and b) the model sampling (Δ_{mod}) methods for the summer surface air temperature over North America for three time periods (from left to right) : 2000-2020, 2040-2060 and 2080-2100 relatively to the 1900-1950 period. All the available simulations are used in the computation.

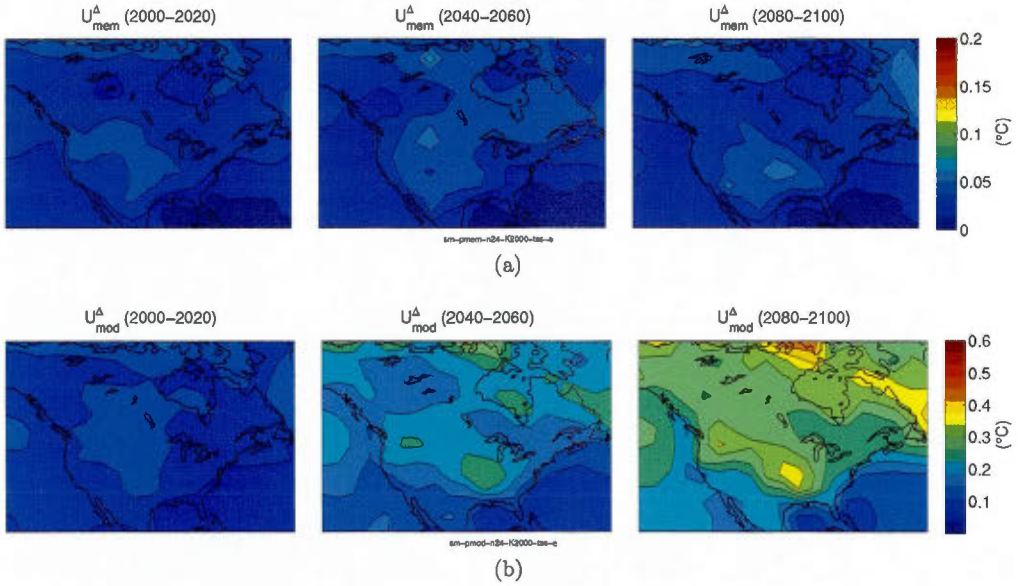


Fig. 1.2: Uncertainty of the signal mean value due to a) the member sampling (U^{Δ}_{mem}) and b) the model sampling (U^{Δ}_{mod}). All the available simulations are used in the computation.

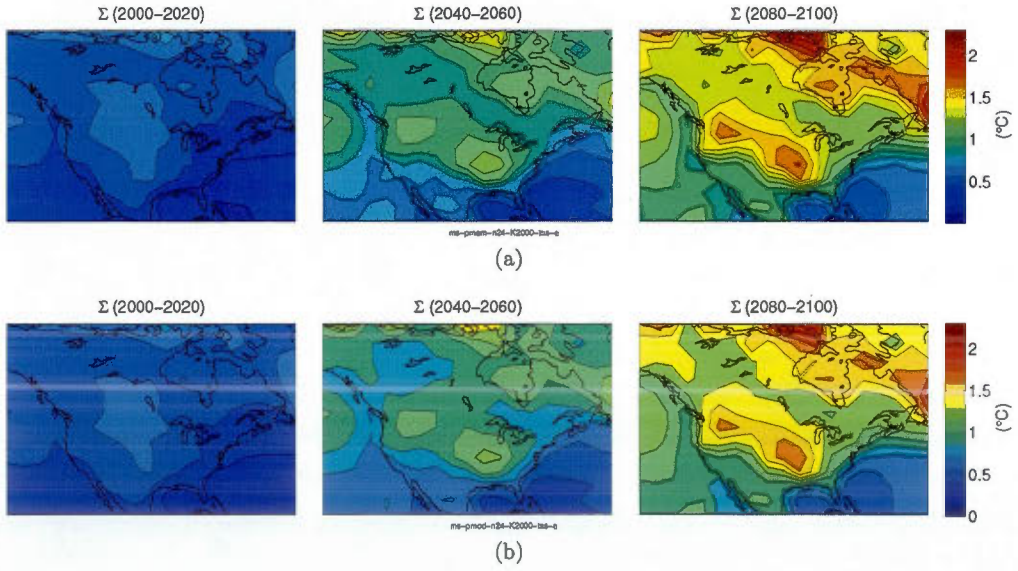


Fig. 1.3: Inter-model spread mean value calculated using a) the member sampling (Σ_{mem}) and b) the model sampling (Σ_{mod}). All the available simulations are used in the computation.

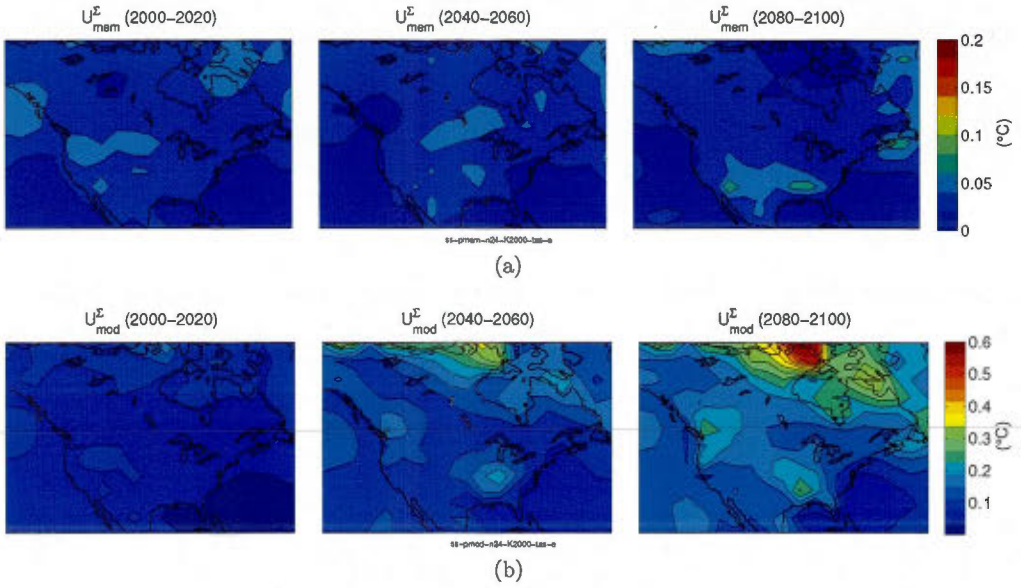


Fig. 1.4: Uncertainty of the inter-model spread mean value due to a) the member sampling (U^{Σ}_{mem}) and b) the model sampling (U^{Σ}_{mod}). All the available simulations are used in the computation.

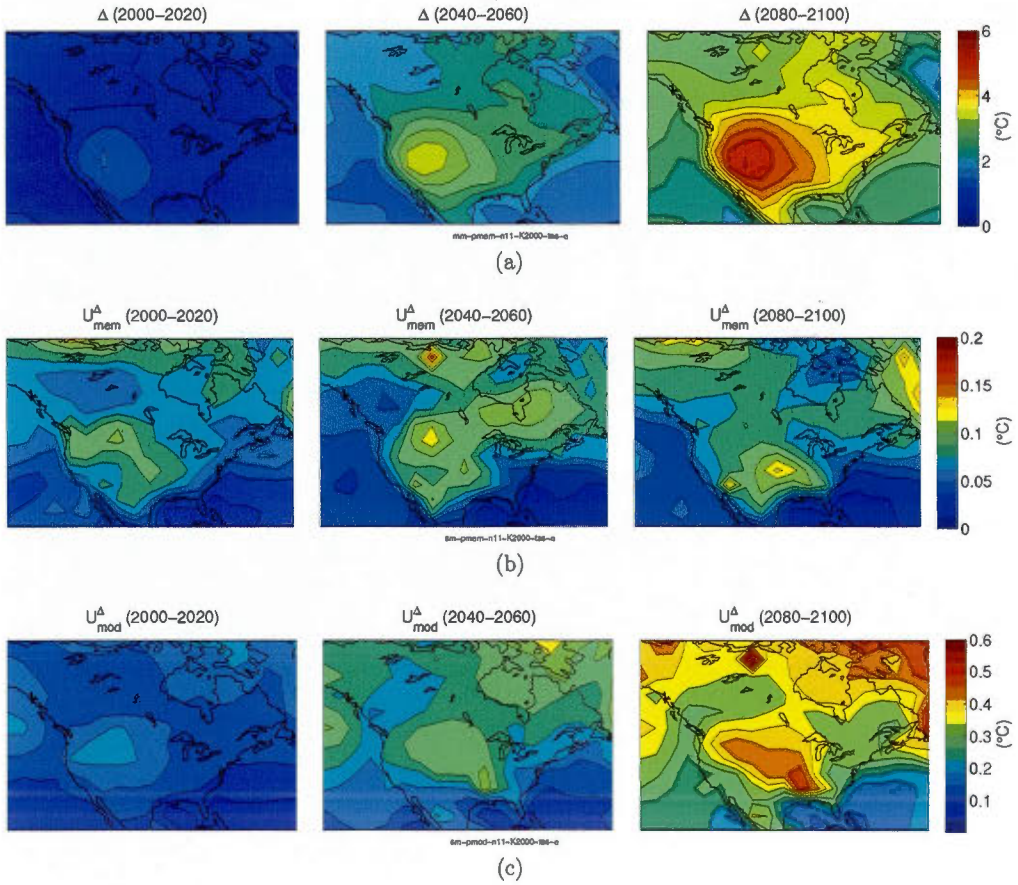


Fig. 1.5: a) Signal mean value (Δ) and its components of uncertainty due to b) the member sampling (U^{Δ}_{mem}) and c) the model sampling (U^{Δ}_{mod}), calculated using the 11-model subset.

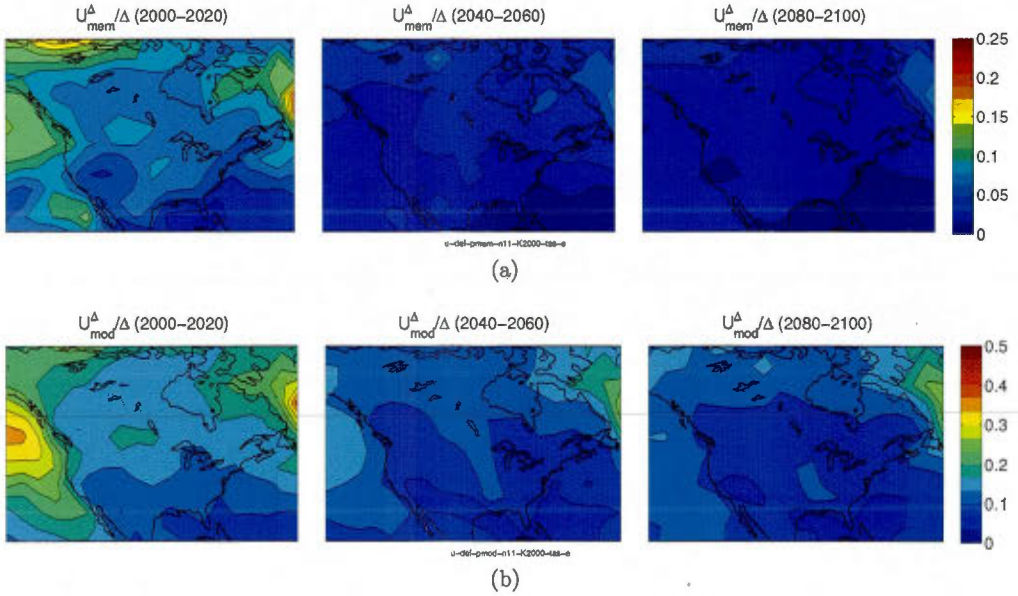


Fig. 1.6: Relative uncertainty of the signal mean value due to a) the member sampling (U^{Δ}_{mem}/Δ) and b) the model sampling (U^{Δ}_{mod}/Δ), calculated using the 11-model subset.

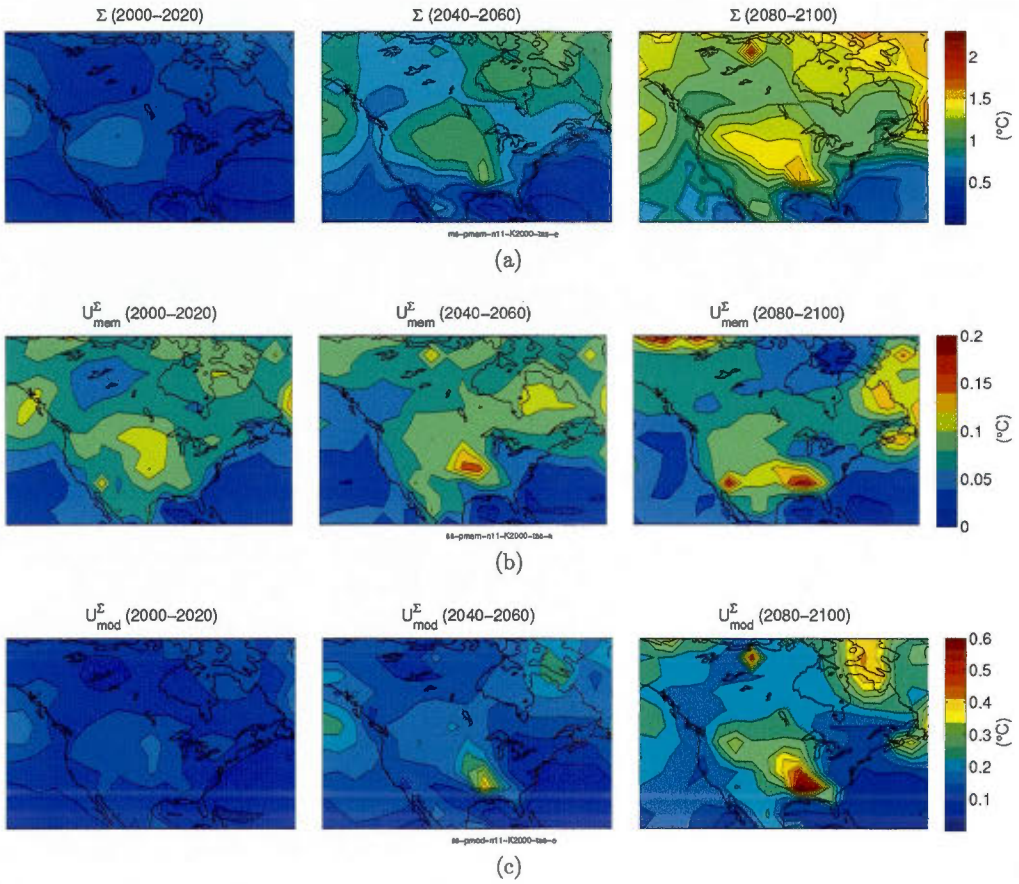


Fig. 1.7: a) Inter-model spread mean value (Σ) and its components of uncertainty due to b) the member sampling (U^{Σ}_{mem}) and c) the model sampling (U^{Σ}_{mod}), calculated using the 11-model subset.

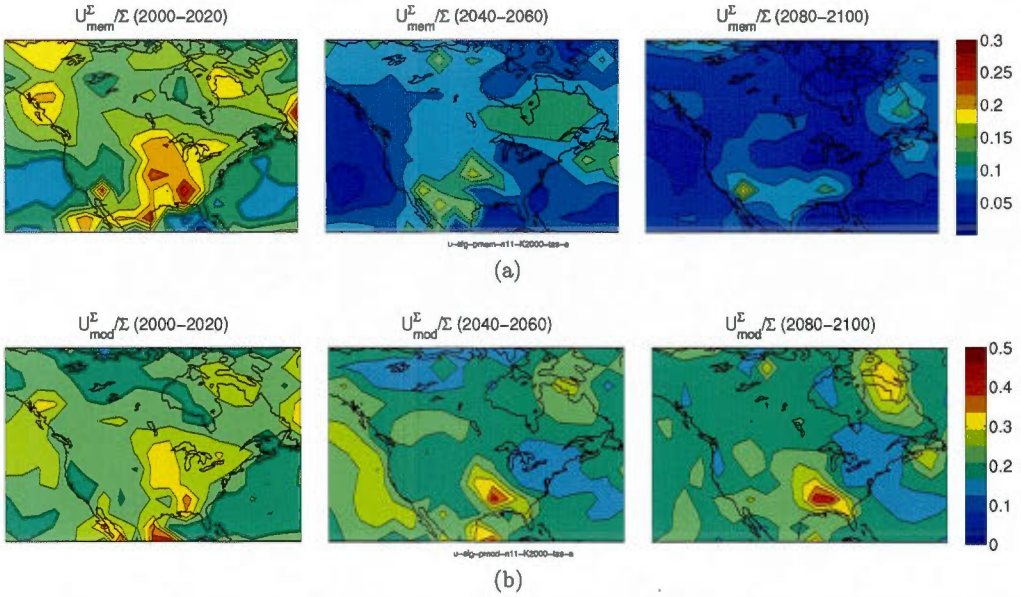


Fig. 1.8: Relative uncertainty of the inter-model spread mean value due to a) the member sampling (U^{Σ}_{mem}/Σ) and b) the model sampling (U^{Σ}_{mod}/Σ), calculated using the 11-model subset.

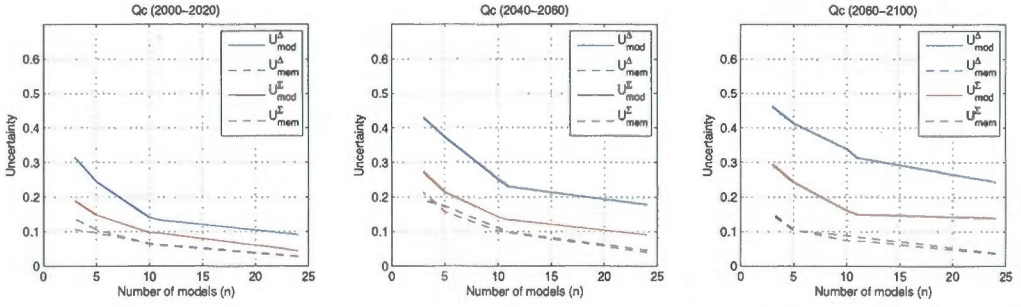


Fig. 1.9: Uncertainty components for the signal and the spread as function of the number of models in the ensemble for a grid point located at the centre of the Québec province of Canada.

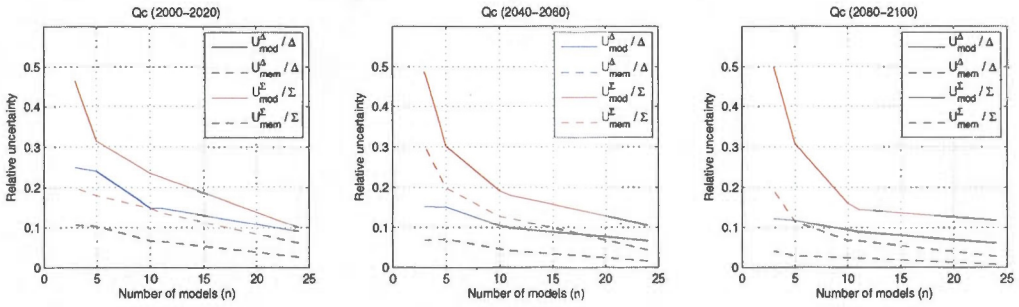
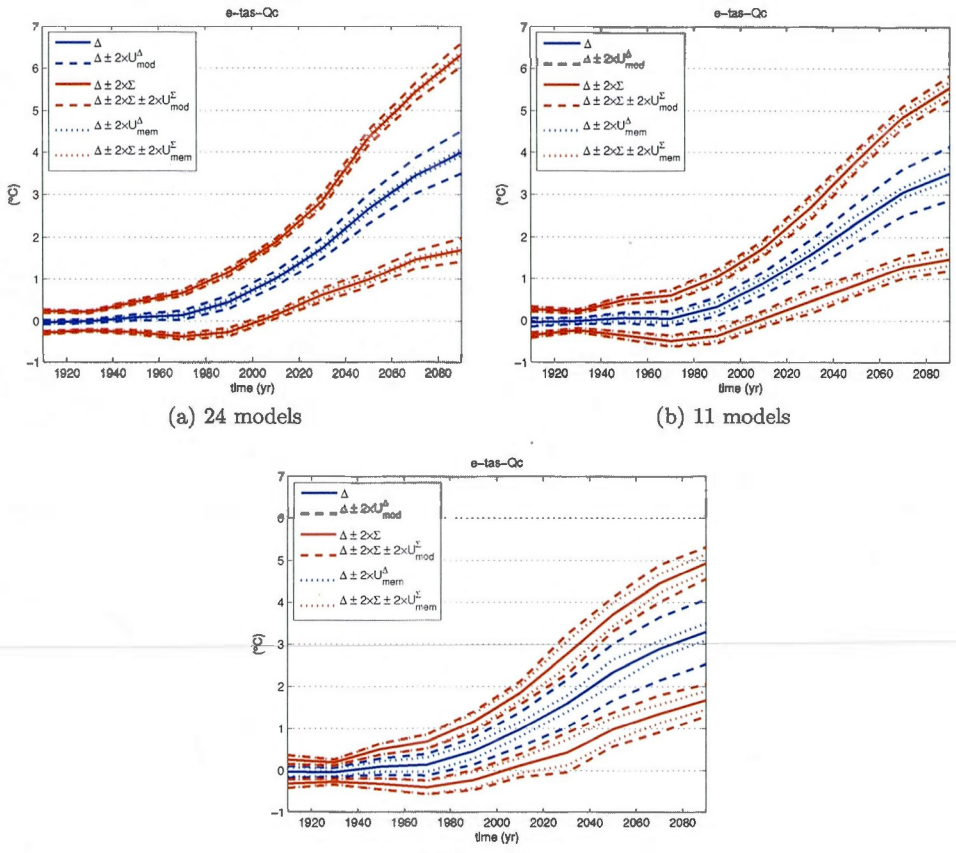


Fig. 1.10: Relative uncertainty components for the signal and the spread as function of the number of models in the ensemble for a grid point located at the centre of the Québec province of Canada.

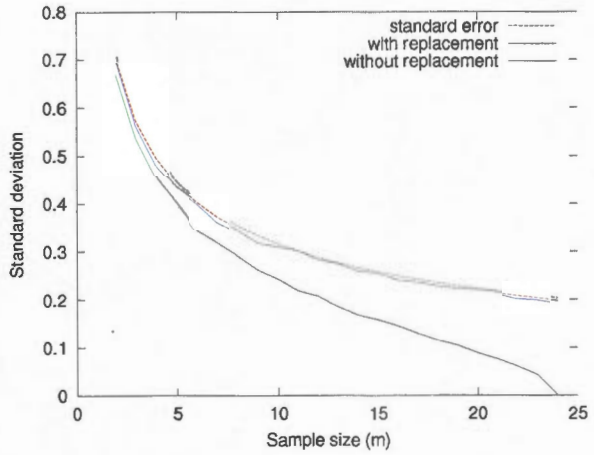


(a) 24 models

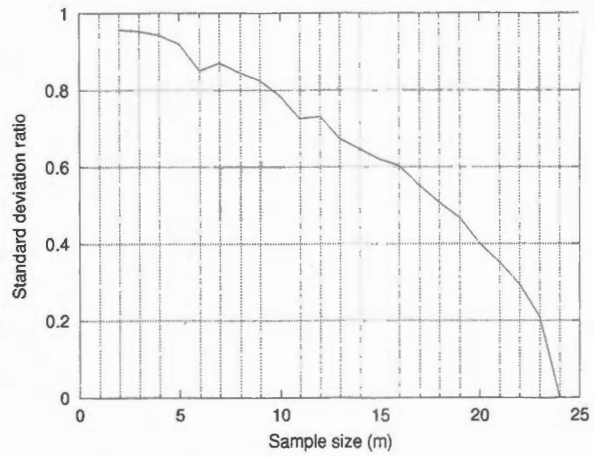
(b) 11 models

(c) 5 models

Fig. 1.11: Plume diagram for the surface air temperature in the summer season over a grid point centred over the Québec province of Canada. The blue and red full lines consist in the signal and inter-model spread mean values respectively, the blue and red dashed lines are the statistical uncertainty of the signal and inter-model spread mean values using the model sampling method, and the dotted lines the statistical uncertainties using the member sampling method. The plumes are obtained from three different ensemble sizes : a) the entire 24-model ensemble and the b) 11-model and c) 5-model subsets.

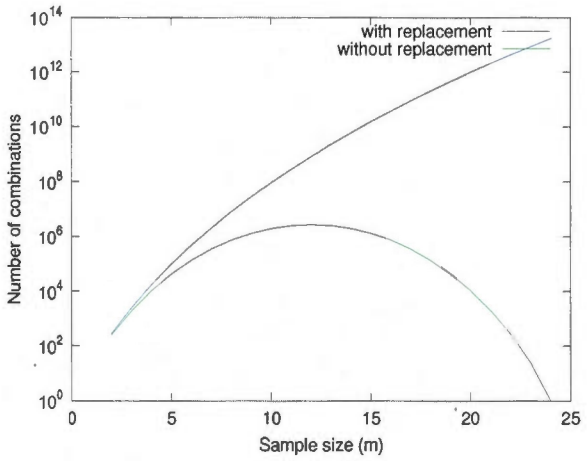


(a)

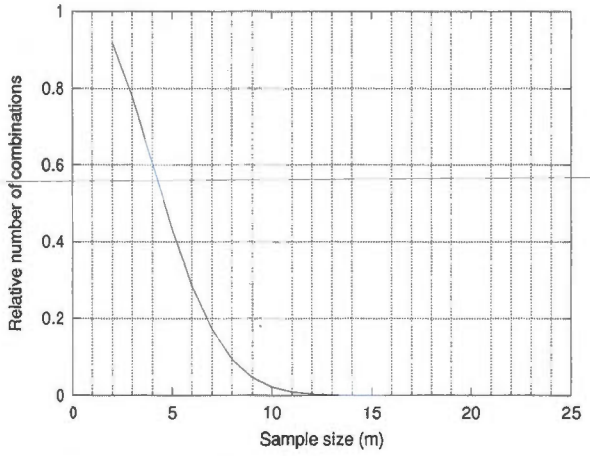


(b)

Fig. 1.12: a) The standard deviation of the mean as function of the sample size obtained from a synthetic data set generated using a random number generator based on a normal distribution with zero mean and unit variance. The initial data set consists in 24 elements over which is applied the model-sampling approach by allowing and forbidding model replacement (blue and green curves respectively). The curves are normalised using the standard deviation of the initial data set and compared with the normalised standard error relationship (in red) defined as $1/\sqrt{m}$. b) The ratio of the errors given by the green and blue curves in (a).



(a)



(b)

Fig. 1.13: a) The number of combinations that can be formed from an ensemble of 24 models as function of the sample size. In green is shown the number of the combinations that can be formed without replacement. The blue curve represents the total number of combinations, including both with and without replacement possibilities. The blue curve is based on the fact that $\binom{N+n-1}{m}$ multisets of size m can be formed from a pool of N elements while the green curve represents the $\binom{N}{m}$ possible subsets. b) The ratio of the numbers given by the green and blue curves in (a).

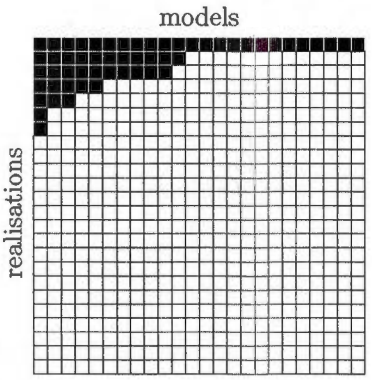


Fig. 1.14: The "MME mask" where the black elements ("TRUE" values in the code) represent the CMIP3 simulations using the A1B scenario and the white elements ("FALSE") stand for the missing simulations in the ensemble compared to the perfect matrix P. Models are distributed along the horizontal axis and the members along the vertical one.

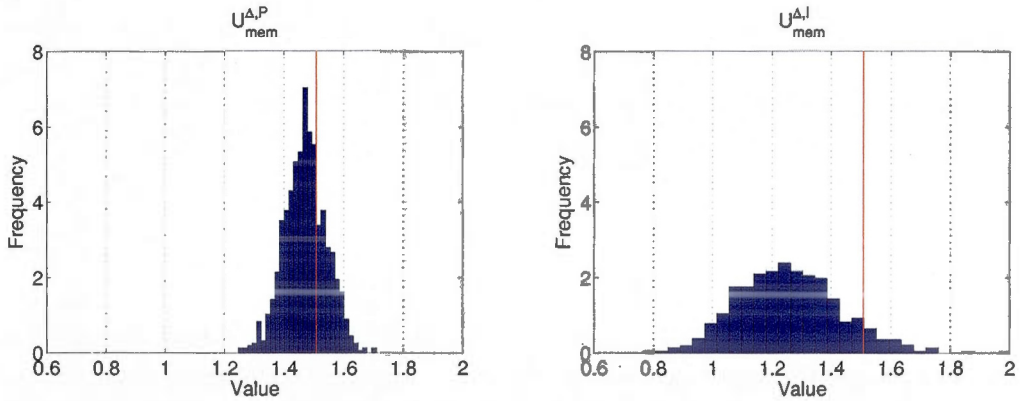


Fig. 1.15: Distribution of the uncertainty emerging from the member-sampling approach for the perfect ($U^{\Delta,P}_{mem}$, left panel) and imperfect ($U^{\Delta,I}_{mem}$, right panel) matrices. Frequencies are normalised to obtain an integral of 1 under each distribution.

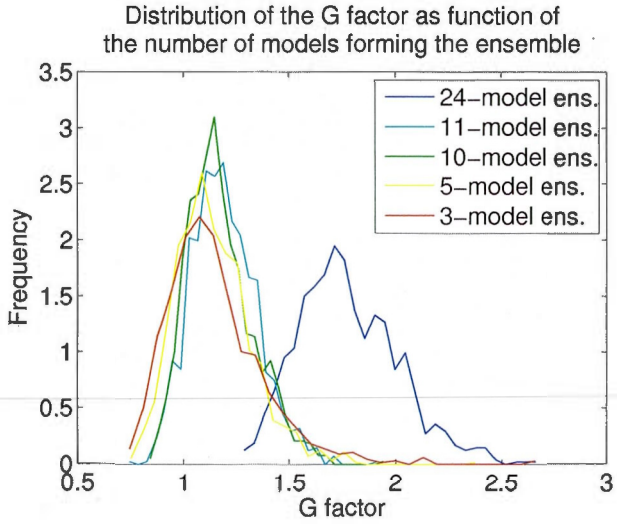


Fig. 1.16: Distributions of the bias-correction factor (G) for ensembles of 24, 11,10, 5 and 3 models. Frequencies are normalised to obtain an integral of 1 under each distribution.

CHAPTER II

INVESTIGATING CONSENSUSES IN CLIMATE-CHANGE PROJECTIONS FOR MODELS DEVELOPED BY A SAME RESEARCH INSTITUTE

ABSTRACT

One rationale behind the use of multi-model ensembles is the aim at collecting independent estimates of the future climate change. Such projections are generally provided by different leading modelling centres around the world, resulting in ensembles that are intended to allow some cancellation of errors as their sample size is increased.

In theory, two unconnected groups of scientists could be expected to develop independent modelling approaches. However, there are in reality several reasons to question this assumption of independence. For instance, scientists share knowledge about the climate system, which is likely to result in models that are based on similar sets of physical assumptions in their formulation. Some models even share parts of their code and are often calibrated using similar observational datasets. All these facts contribute to the risk of inducing common biases to the models and hence to a lack of independence.

While a rather classical approach to assessing model independence could consist in detecting possible correlations of errors when comparing models to the observations, such an approach can not be directly applied to climate-change simulations due to the relatively short climate period available for validation. An alternative approach can be to investigate directly on the differences between the models' climate-change projections. An additional issue resides in determining whether if similarities in the models output consist in a proxy of high confidence into a specific climatic outcome or simply due to a lack of independence between the underlying models. In order to improve the message conveyed by an ensemble, it is of primary importance to aim at filtering the non-informative consensuses from the ensemble in order to focus only on the informative ones. On the other hand, disagreement between models' output could be seen as informative from the point of view of assessing the uncertainty related to the use of different coexisting modelling approaches.

2.1 Introduction

In the last decades, internationally coordinated efforts have been conducted in order to nourish the scientific community with credible ranges of climate-change projections. These projects consist in relatively large ensembles of simulations that aim at sampling the different sources of uncertainty affecting climate-change projections. Firstly, the emission scenarios of greenhouse gases and aerosols (GHGA) used as an external forcing to the models depends on the evolving socio-economical context and hence play an important role in our uncertainty of the future climate changes. Secondly, the use of a population of state-of-the-art Atmosphere-Ocean General Circulation Models (AOGCMs) allows obtaining a considerable range of projections since models generally differ in their climate sensitivity for a given forcing scenario. Thirdly, each model/scenario combinations are often subject to several members (realisations) that differ only by their initial conditions; multiple realisations allow sampling the natural variability as simulated by the models, which is also considered as a source of uncertainty affecting climate-change projections.

It is a generally accepted idea that climate models are not "created equal", that is they perform differently in reproducing the various facets of the climate system (Gleckler et al., 2008). An important reason why climate models perform differently is that they are based on different approximations, which are sometime subject to debates within the climate modelling community. Given that a model may perform well in reproducing some climatic features while showing weaknesses in simulating others, it has been suggested that climate-change projections from several models should be weighted according to some measure of their respective skill. As an example, Giorgi and Mearns (2002) developed a method for obtaining weighted averages based on both the performance of the models in reproducing the observed climate and the consensus of their projections of the future climate. Similarly, Christensen et al. (2010) used several metrics for evaluating models' performance and combined these scores into a single weight for each model. The best way for evaluating models' performance and to use this measure of skill

for assigning weights to the models is however far from making any consensus among the climate scientists. Moreover, it is poorly understood how the skill of a model in reproducing the present climate may be related to its reliability for climate projections (Räisänen, 2007).

Nevertheless, the rationale behind the use of multi-model ensembles is to collect several independent estimates of future climate changes. Under the assumption of independence, a cancellation of errors is expected to happen between the different estimates and hence the ensemble average should be closer to the climatic truth than any single model. Another expected advantage of using independent estimates of climate changes is that the spread of the members in an ensemble provide an estimate of the uncertainty about the modelled system. There are prior reasons, however, to believe that models are not totally independent from one another. While models differ in a variety of facets since several expert's choices are involved in the development of such complex softwares, climate scientists share knowledge, learn from each other and even share parts of their model code. This results in climate models that are similar at different levels, from their underlying physical assumption (included processes and interactions), the tuning of weakly constrained internal parameters, and the numerical approximations used to solve the equations.

It is a generally accepted idea that climate models (and hence their projections) suffer from a lack of independence, but how to determine to what extent ? Very little is known for answering this question since no clear metric for measuring model independence has been commonly accepted at this time (Tebaldi and Knutti, 2007). Fortunately, some attempts have been made over the last years for assessing the degree of independence of the climate models, which are mainly based on three points of view.

A first way of addressing the independence of climate-change projections is by focusing on the models formulation, i.e. *a priori* to generating output. Probably the largest issue related to this approach is the definition of a model space, which can not be done using real numbers by analogy to perturbed-physics ensembles (PPEs ;Stainforth et al. 2005;

Rowlands et al. 2012) that offer a systematic approach to exploring the parameter space of a single model.

Another approach for assessing the independence of climate models simulations is by considering that independent estimates are evenly distributed about the truth, that is the observed climate. This approach has been used by some authors (Jun et al. 2008b,a; Knutti 2010; Pennell and Reichler 2011) in the context of the Coupled Model Intercomparison Project phase 3 (CMIP3). A general result from these studies is that partial correlations exist between the models' biases to the observed climate and hence that the models' output is generally not evenly distributed about the climatic truth. However, a disadvantage of such an approach is that it can not be applied directly to climate-change projections due to the relatively short climate period available for validating the models and since no observations are available on the future state of the climate system. In order to make an inference about the independence of the climate-change projections, one has to assume that a sample of models providing independent estimates of the observed climate will necessarily lead to independent projections of the future climate.

The third approach addresses the issue of independence from the point of view of the dissimilarities in the models output. A clear advantage of such an approach is that it does not need an observational data set and hence that it can be applied to climate-change projections. Abramowitz and Gupta (2008) projected the model space onto a metric space from which the distance between two models can be used as a proxy for model independence. They also put in evidence that model independence and model performance consist in two unrelated properties of climate models (Abramowitz, 2010). More recently, Masson and Knutti (2011) used a hierarchical clustering framework according to the degree of similarity in the models' projections and put in evidence that models developed by a same institute are likely to provide similar results.

We will proceed with an overview of the issues related with the use of multi-model ensembles for climate-change assessment. In Sect. 2.2 we will analyse the sampling process related to an ensembles of opportunity. In Sect. 2.3, we will review and discuss the

topic of model performance, following with a theoretical discussion about model independence in Sect. 2.4 by focusing on the conceptual relationship that may exist between prior considerations about the models (independence of the models formulations) and the consensus/disagreement in their output (independence of the output). We will analyse in Sect. 2.5 the typical structural similarities that appear between models developed by a same research institute and present some results based on a subset of the CMIP3 multi-model dataset. In Sect. 2.6, we note some lacks that have been found in the documentation provided by the participating centres about their models and simulations. We finally proceed to a broader discussion (Sect. 2.7) about model independence and focus on the possible ways to improve our interpretation of an ensemble of opportunity according to the way these are constructed and used by the scientific community.

2.2 On the sampling process of an ensemble of opportunity

The Program for Climate Model Diagnosis and Intercomparison (PCMDI) initiated the sampling process of the CMIP3 multi-model dataset by volunteering established modelling centres to participate by delivering AOGCM climate projections. In order to participate to such a coordinated experimental framework, the modelling centres are generally committed to some minimal requirements, for example by delivering the simulations before some deadline and according to a specific data format. Requirements on submitted variables are rated as low to high priority in order to focus on specific scientific issues. Finally, a variety of experiments have been proposed to the modelling groups, the main ones being the 20th century experiment (labeled as 20C3M) and the projections using three emission pathways (A1B, A2 and B1) from the special report on emissions scenarios (SRES; Nakicenovic et al. 2000).

In order to encourage (and hence maximise) the amount and diversity of simulations in the ensemble according to the number of scenarios, models and realisations, few other constraints are imposed to the participating centres. It is indeed very expensive to produce simulations for long periods (e.g. centuries) with increasingly high spatial reso-

lutions. Because human, computational and funding resources are limited and modelling centres do not all share the same interests, the final size and shape of the ensemble is necessarily affected by these factors. Hence an arbitrary numbers of realisations are generated using the different models, some institutes may also provide simulations from several models or versions, and not all emission scenarios are used to force each of the participating models.

An important particularity of such ensembles is that they do sample different modelling approaches, but in a neither a random or systematic way (Knutti, 2010). The sampling process of such an ensemble could be more akin to a "natural pre-selection" among modelling approaches, where some centres may not afford all of the proposed experiments while others tend to be better represented in the ensemble. One possible drawback of this pre-selection process is that it tends to give a larger "ideological weight" in the ensemble to the better endowed institutes according to the number of provided simulations and their diversity in representing the several scenarios, models, versions and realisations.

2.3 Performance of climate models

Climate-change assessments face several issues when attempting to extract the message conveyed by an ensembles of opportunity through the use of ensemble statistics. An important issue that is under debate in the community is the optimal way of combining simulations from different models. Probably the simplest and easiest manner to process multi-model ensembles is known as "one model, one vote" (Knutti et al., 2010), which considers the models as equivalent representations of the climate system. This can be interpreted as assuming that each model is independent from one another and hence that the models' simulations consist in equally likely outcomes of the future climate. The ensemble mean is interpreted as a best estimate of the projected signal since individual model errors are expected to cancel out through ensemble averaging. In addition to its simplicity, this approach is widely used since the average of several models often outperforms each of the individual models of an ensemble in reproducing current climate

(Lambert and Boer 2001; Weigel et al. 2008; Gleckler et al. 2008; Reifen and Toumi 2009; Annan and Hargreaves 2011).

An alternative approach for combining simulations from a multi-model ensemble is based on the assertion that "models are not created equal", i.e. that some perform better that the others in reproducing the observed climate, which is taken to imply similarly for the projected climate changes. Indeed, models have different strengths and weaknesses in reproducing the various facets of the current climate (e.g. different variables and geographical locations). If there were a commonly agreed measure of model skill by the scientific community, a convenient way to process multi-model ensembles would be to give preponderance to the "best models" while down-weighting the "bad ones" in ensemble averaging. However, since a very short period of observation is available for model verification, and the reliability of climate models in projecting future climate changes can not be assessed directly, it is not clear which climate features have to be better reproduced in order to increase our confidence in model projections. As an example, Christensen et al. (2010) defined six metrics of model performance based on the skill in reproducing the annual cycle, trends, large-scale circulation, etc. These metrics have been combined in order to assign a single weight to each of the models in calculating the ensemble statistics. Another example is Giorgi and Mearns (2002) who used the model performance in reproducing the observed climate and the consensus between the model projections in order to obtain the ensemble mean and standard deviation that are weighted according to these two criteria; the authors however noted that the performance of the models in reproducing the observed climate is poorly related to their consensus in climate-change projections.

While evaluating the model performance is far from trivial due to the paucity of climatic data for verification, another important characteristic of a multi-model ensemble is the degree of independence that exist between research institutes and modelling approaches. As will be explained in the next section, there are several indications that the models forming the CMIP3 multi-model dataset suffer of a lack of independence. Unresolved issues

in attempting to obtain a clear measure of model independence limits our interpretation of ensemble statistics.

2.4 Independence of climate models

Today's climate models exist in a broad diversity since several expert decisions were involved in the development of these complex pieces of software. Models may differ in a variety of facets including their basic physical assumptions since modellers have to identify and judge which processes of the climate system are sufficiently relevant to be included in a model. For example, in the upcoming CMIP5 multi-model ensemble, a dynamical vegetation component is included for some of the models while others use a static vegetation cover. Another way climate models may differ is how the included processes are formulated, as by choosing among several possible physical parameterizations for a same process (e.g. Bechtold et al. 2001 vs Kain and Fritsch 1990 for convective parameterization). While these two types of model difference can be referred to as "structural", another kind of difference exists between climate models. These differences may appear in the numerical approximations of the equations, the time and spatial resolutions, and the tuning of some poorly constrained parameters (Murphy et al., 2007; Stainforth et al., 2005; Murphy et al., 2004; Separovic et al., 2012). In the following, we refer to this type of difference as "parameters and numerical". It is worth noting that successive versions of a climate model may in principle differ in the same way as models do (i.e. structural, parameters and numerical); however, differences between versions are generally subtler due to limited changes.

Despite the variety of differences existing between climate models formulations, the modelling centres are not completely independent from one another from an ideological point of view, and so are expected to be both the models and their output. As for science in general, the climate science evolves in a rather open manner as scientists share knowledge about the climate system and learn from other groups through literature, conferences and exchanges. An important example of this is the physical basis

of fluid dynamics and thermodynamics that is similarly formulated within the core of every atmosphere or ocean model. To a lesser extent, the research centres also share model components (e.g. parameterization packages) and even parts of code. In particular, this characteristic of the climate models to share common components is likely to be strengthened when the models are developed by nearby actors, for example within a same research centre or country. This may also be due to the fact that several models have public releases that can be downloaded and used by other groups or individuals. Another reason to believe *a priori* that climate models are not independent is because they are often tuned according to the same observational data sets that also contain some errors, a strategy that may induce common biases to the models (Knutti et al., 2010). Moreover, even in the idealised situation where a model would fit perfectly to perfect observational data, it would be possible that a good result be obtained for wrong reasons. While different tunings of the parameters can lead to similar model output, exploring systematically the parameter space of a model would be an humongous task to undertake (Stainforth et al., 2005).

As noted previously, one rationale behind the use of multi-model ensembles is to obtain independent estimates of the future climate changes. However, there is no commonly accepted measure of the degree of independence between climate models (Tebaldi and Knutti, 2007). While several authors use consensus between models as a predictors of confidence (e.g. IPCC 2007; Seager et al. 2007), such an inference is difficult to sustain without any robust measure of models' independence (Pirtle et al., 2010). One important issue is the difficulty to define a model space, which can not be illustrated clearly by using real numbers, for example. By analogy, in a perturbed physics experiment (PPE), the differently tuned versions of a same model can be represented by points in a multi-dimensional space of parameters. The distance between two model versions in the parameter space can be associated to a distance in the projected phase space (model output). Clearly, this approach can not be applied to quantify the model uncertainty since models that differ structurally are also represented by spaces that may differs in both their number of dimensions and in the definition of each axis (parameters). De-

fining a general space that would contain all of the individual model sub-spaces is a conceptual issue that has found no clear answer at this time while being an emerging field of research over the last few years.

Expecting that, *a priori*, climate models suffer from a lack of independence partly questions the way these models are developed and improved over years by scientists around the world. On the other hand, it is also important to understand how, *a posteriori*, these characteristics may affect the models' output. As shown by Knutti et al. (2010), the CMIP3 models are partly correlated in their biases to observations over several regions of the global domain. Other important results have been obtained by Masson and Knutti (2011) who used a hierarchical clustering framework to put in evidence that the degree of similarity between models' output is intimately related to the "model genealogy".

In order to put the previous discussion in the context of climate-change assessment using multi-model ensemble, Fig. 2.1 shows a diagram summarising the conceptual relationships that exist between the prior and posterior considerations of independence. While no widely accepted metric exists for assessing model independence, as well on the side of the models themselves (*a priori*) as in the models' output (*a posteriori*), we posit the following definitions. A set of models are said to be ideologically independent if the modelling approaches differ substantially by their included processes, parameterizations, numerical approximations and tuning of parameters. The two boxes on the left side of Fig. 2.1 represent high and low levels of ideological independence within a sample of models. For simplicity, we assume all the models in the sample to be either independent or not, and to the same extent, unlike the more complex case of the CMIP3 multi-model dataset where some groups of models are more similar than others, corresponding to a mix of different levels of independence. On the right side of the diagram, the two boxes represent two degrees of similarity (disagreement or consensus) that may exist between the models' output. The output can be considered as in disagreement (consensus) if their sensitivity to equal GHGA forcing differs by a larger (smaller) amount than the typical magnitude of the natural variability as simulated by this type of model.

Let us now discuss the conceptual relationships between prior and posterior states of independence. As a first possibility (link A in Fig. 2.1), *a priori* independent models lead to outputs that disagree; this corresponds to the case of a wide range of models' responses to identical climate forcings, a situation that contributes to the so-called model uncertainty. As a second possibility (B), *a priori* independent models lead to outputs that agree; such a consensus between substantially different modelling approaches generally tends to reinforces our confidence into a specific outcome in the simulation of the future climate (so-called robust results). The third relationship (C) shown in the diagram relates to the trivial situation where the sample is formed by models being very similar in their structure, parameters and numerical characteristics; such a case consists in a "non-informative consensus" since based on several replications of (essentially) a same model. The fourth hypothetical case (dashed line) of *a priori* non-independent models leading to disagreeing outputs is obviously unrealistic since a sample of models that are considered as replications of a single one should not lead to differences in their outputs that are larger than the simulated natural variability, unless some modelling differences are hidden to the data user which then should be interpreted as case A.

In summary, the analysis of the climate projections obtained from an ensemble of opportunity would be highly simplified if we could assert that only relationships A and B exist, which would clarify the meaning of consensus as the most likely outcome and inter-model spread as a measure of uncertainty in the projections. We argue that disagreement between models' output is always informative, unless serious bugs are known to exist in some of the models. From the point of view of mitigation and adaptation strategies, it is generally more cautious to deal with overestimated uncertainty in order to assume a larger range of possibilities for the future climatic outcome. On the other hand, underestimated ranges of uncertainty simplify the mitigation process while increasing the risk due to an unsuspected, damaging and costly climate outcome to happen.

It is worth noting that the previous discussion focuses on multi-model ensembles but might apply as well to PPEs that face similar issues. For example, more or less inde-

pendent attributions of the model parameters may be fetched through these experimental frameworks, which are however constrained according to a unique model structure.

2.5 Typical differences between models developed by an institute and how this affects their climate-change projections

AOGCMs constitute the main tools used by scientists in order to better understand the present climate and its projections to the future according to conceivable GHGA pathways. The diversity of modelling approaches resulting from the various choices available for structural, numerical and parameter characteristics, makes a complex task to describe the differences existing between two models that have been developed independently. In order to simplify the following discussion, we focus on the typical differences among models that share a considerable number of components, i.e. with a certain level of structural similarity. By analogy to the model space briefly described in Sect. 2.4, it consists in comparing models that belong to partially different parameter spaces. Rather than proceeding to an exhaustive study of all the differences in structural, parameters and numerical characteristics of some twenty AOGCM models in the CMIP3 multi-model ensemble, an intuitive approach to identify models with structural similarities is paying attention to their origin. As an educated guess or proxy for model non-independence, the models being developed by a research institute can be expected to share several characteristics.

In Tab. 2.1 are presented the 7 research groups (first column), hosted by 5 countries (second column), that provide more than a single model to the CMIP3 multi-model archive. In the third column is given an acronym that represents the research group and their models. In Tab. 2.2 is shown the list of the corresponding models collapsed into pairs developed by a same research group. The pairs of models are numbered from I to IX (first column) and identified by their acronym in the second column. In the third column of Tab. 2.2 are shown the models (or versions) identifiers. The five remaining columns in Tab. 2.2 enumerate the main structural differences according to main model

components : atmosphere (A), ocean (O), sea ice (I), coupling (C) and land surface (L). Each unit in these columns is filled with an identifiers for the type of difference : R for resolution, V for version (i.e. minor modifications in the code), M for model (i.e. substantial differences in the code), and "-" when the same component is used. This table consists in an adaptation from Randall et al. (2007) and more details about models characteristics are given in the PCMDI documentation at `http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php`.

According to Tab. 2.2, the pairs I to IV are formed by models that differ little in a structural sense. The CGCM models (I) differ only in the atmosphere and ocean spatial resolution (T47 vs T63), and similarly for the MIROC pair (II) with a larger jump in resolution (T42 vs T106). Changes in resolution only could be considered as a parameters modification while no other changes are expected in the code. The CSIRO (III) and GFDL (IV) pairs provide models' versions that differ in minor modifications to their main components. More precisely, the version change may apply to any of the model components, i.e. atmosphere and ocean for GFDL, and ocean, ice, land and coupling for CSIRO.

The pairs V to IX are formed by models differing substantially in a structural sense. The first GISS pair (V) consists in two models (EH and ER) that differ in the ocean component only. Rather than successive versions, two different ocean models (Russell et al. 1995; Bleck 2002) have been used in these AOGCMs, which could be seen as a substantial structural difference. In addition to this structural difference, the two ocean components used different spatial resolutions. For the pairs VI to IX, models differ substantially according to most of their components (atmosphere, ocean, sea ice, coupling and land). An apparent similarity however exists between the models AOM and ER (in pair VI) which use successive versions of the same ocean model (Russell et al., 1995, 2000). Among other differences, each of the pairs VI and VII is subject to a different atmosphere component. Overall, GISS AOM, EH and ER can be seen as coexisting models developed within a same institute (NASA/GISS) but one (AOM) appears more

different from the two others (EH and ER). Similarly, pairs VIII and IV are formed by two coexisting models developed within a same institute, i.e. the National Centre for Atmospheric Research (NCAR) and the Hadley Centre.

In Tab. 2.2 have been presented the main structural differences that appear within 9 pairs formed by 15 models and developed by 7 research institutes. In Fig. 2.2 are summarised the results from these models, according to their domain averaged (North America) and ensemble averaged (the mean of all the available realisations of each model) for the surface air temperature and precipitation rate for the summer and winter seasons. The research institutions are represented by different colors and the models from a same institute are identified using different line styles. We first note the high consensus between the two versions of the Canadian model that use different spatial resolutions (dark blue curves). The two versions strongly agree with a change of approximately 3.4°C for the summer temperature warming (Fig. 2.2a) and agree relatively well in winter (Fig. 2.2b) with a temperature increase of 5.8°C. Even though noisy, the summer precipitation rate also shows relatively high agreement between the two model versions while a slight difference is found for the winter season. By comparison, the MIROC pair (in magenta) displays rather large differences in sensitivities with exception of the summer temperature where the models agree relatively well with a change of 5.3°C. The disagreements are more important than those between the two versions of the Canadian model, probably due to the larger increase in resolution (T47 to T63 for CGCM, while T43 to T120 for MIROC). Moreover, this result could also suggest that modifications to other parameters have been done between the two MIROC models, but little information has been found in the available models' documentation.

Another striking feature in Fig. 2.2 is the pair V formed by the GISS EH and ER models that differ only by their ocean model (i.e. same atmospheric, land, coupling and sea ice components). These two models agree generally well according to the two variables and two seasons presented in the figure. Also, the two coexisting models of the NCAR institute (pair VIII) show very similar results for precipitation (Fig. 2.2c and d).

For the Hadley Centre that also provides two models, some consensus is obtained for the summer temperature and winter precipitation (Fig. 2.2a and d), while the two other variables (Fig. 2.2b and c) show disagreement that appear similar in magnitude to the overall inter-model spread.

While the inter-model spread is a measure of uncertainty that highly depends on the selected set of models, it is convenient to compare the differences in the models' output according to the natural variability that constitutes an intrinsic characteristic of the climate system. However, the main limitation for such an approach resides in the number of realisations available for each of the models. Over the 9 pairs of models shown in Tab. 2.2, three pairs (CSIR0, GFDL and UKMO) are formed by models with a single realisation and hence can not be used in the scope of the following analysis. In Tab. 2.3 are shown the number of members available for each model in each pair (N_x and N_y) and the fifth column shows the feasibility of a Student's t-test for the difference between ensemble means.

The difference of mean between each model pair is therefore calculated. We focus on the summer surface air temperature change relative to the 1900-1950 reference period, and all models have been interpolated over a common $4° \times 5°$ coarse-resolution grid (see Sect. 1.2.1 for more details). In order to assess the statistical significance of these differences, a two-tailed Student's t-test is applied at the 5% significance level (i.e. 2.5% on each tail). As will be discussed in Chap. 3, the models show differences in their simulated natural variability. For pairs where both of the models provide at least two members, the t-test can be applied without assuming equal variances. On the other hand, the pairs where one model is represented by a single member are restricted to the assumption of equal variances. More details on the t-test are provided in Appendix 2.A.

In Fig. 2.3a to f are presented the differences of the ensemble means between models developed by a same institute, for the pairs I, II, V, VI, VII and VIII respectively. The order in which the differences are calculated corresponds to that presented in Tab. 2.2, for example as T47 minus T63 for the CGCM pair. Each panel is composed of six maps,

where three time periods are shown : 2000-2020, 2040-2060 and 2080-2100. On the first row are shown the differences in ensemble means and in the second the areas of statistical significance for positive (red) and negative (blue) differences. White areas corresponds to regions where the differences are not statistically significant.

For the CGCM and MIROC pairs, (Fig. 2.3a and b), an important part of the domain shows differences that are not statistically significant. For both pairs, a similar pattern in the rejection of the null hypothesis is seen over the Atlantic Ocean (south eastern part of the domain) and in the region of the Labrador Sea and Hudson Bay (eastern side of the domain). It is worth noting that the opposite sign of these patterns is simply due to the order of the difference is calculated, and hence the higher resolution model leads to a larger climate sensitivity for both pairs. It is worth noting that the two model versions have been interpolated over the same grid and hence an important part of the potentially added value by the higher resolution models is not considered here. On the other hand, it can be seen that the change in resolution has a rather weak effect on the larger scales present on this grid. According to the small sample size, the coarse resolution and the variable considered, there is little statistical evidence to reject the hypothesis of an equality of the means between these model versions. Under these considerations, including both model versions for the CGCM and CSIRO pairs does not add much supplementary information to the ensemble compared to the use of a single model version.

In Fig. 2.3c is shown the difference between the GISS EH and ER models that differ in their ocean component model. Significant positive differences are found over the Hudson Bay and the maximum increases in magnitude with time to reach nearly 3°C in the 2080-2100 period. Other significant differences are seen over the Pacific Ocean, but these are generally smaller in magnitude than -1°C. From this panel, it is quite clear that structural differences in the ocean component affect mainly the results over oceanic regions and particularly the Hudson Bay. In Fig. 2.3d are shown the differences between the GISS AOM and ER models that have large structural differences in all of their

components except the ocean that underwent only a change in version. An important difference that exceed $-2°C$ appears over land. This difference grows with time, which means that the models' climates diverge from each other over this region, similarly to what has been noted for the EH and ER models over the Hudson Bay (Fig. 2.3c). Differences over the Pacific Ocean have similar magnitude, with values around 1°C. Also, no maximum difference appear over the Hudson Bay, which can be attributed to small structural changes between the two versions of the ocean component. The third pair of the GISS institute, AOM vs EH, is shown in Fig. 2.3e. This time, all of the models' main components have been changed. It is interesting to note the few similarities with the previous GISS pairs differences. The large difference over the Hudson Bay is similar in magnitude with $> 2°C$ (with reverse sign) compared to the EH-ER difference. This response is expected since the AOM-EH pair relates the difference between the Russell (second version) and HYCOM ocean components, while the EH-ER pair corresponds to a difference between the HYCOM and Russell (first version) components. As seen from the small difference between AOM and ER over the Hudson Bay, the two versions of the Russell ocean component do not lead to large differences especially in that region. For the minimum of difference that exceed $-2°C$ over land, it is nearly the same value as for the AOM-ER difference presented in Fig. 2.3d. Indeed, the pairs AOM-EH and AOM-ER (Fig. 2.3d and e respectively) correspond to the same differences in the atmosphere component. This logic may also be applied to the differences over the Pacific Ocean. The difference between ER and EH being around 1°C degree (Fig. 2.3c) and that between AOM and ER of approximately 1°C (Fig. 2.3d), it is understandable how the difference between AOM and EH may be of nearly 2°C (Fig. 2.3e). It is interesting to note that the analysis of the differences between the three GISS pairs leads to differences in their climate sensitivities that are additive from the point of view of structural changes in their atmosphere and ocean models components.

Finally, the difference between the two NCAR models is shown in Fig. 2.3f. These models show statistically significant differences over practically all of the North America. These differences are relatively large and increase with time to exceed 2°C over a large part of

the domain for the period 2080-2100.

2.6 Notes on the minimal requirements to the participating centres of a climate change assessment

In order to facilitate the analysis of the results from a large ensemble such as the CMIP3 multi-model dataset, the participating centres have been invited to fill a survey that summarise their experimental set-up by providing information about model identity, component model characteristics and simulations details. We acknowledge that such pieces of information have an important value to the data users such as ourselves here. Notable lacks exist however which are worth mentioning for the benefit of further users and assessments.

While it is mainly the responsibility of each modelling centre to provide complete and accurate information throughout these surveys, it is important to note that no (or little) control seems to have been applied after their submission to PCMDI. As a proof by contradiction, a minimal post-control on these surveys would not have resulted in the following examples of inaccurate or even missing information. A first example is found in the documentation for the CCSM3 model that contains entries such as "Still working on this..." or "See the excel chart [...] that I mailed you last week" in the section of the simulations details. Another striking example is the missing of such surveys for both model versions from the Canadian Centre for Climate Modelling and Analysis (CCCMA).

One reason why such surveys are very important from the point of view of the data users is since the models configuration often rapidly changes with time and such modifications are not always clearly documented throughout peer-reviewed literature. Another note about these surveys is that the community would largely benefit from a broader focus on the modelling differences between models (or versions) provided by a centre. For example, a centre could be invited through the survey to provide arguments describing how a model or version may add supplementary information to the ensemble, specifically

in the context of collecting independent estimates of the future climate changes and in the aim at spanning the full range of the uncertainty about these projections.

Of course, modelling centres are always welcome to provide several models versions, which may be used in a variety of applications. However, it is worth questioning that the resources spent in these supplementary models or versions could be relocated for example by producing more scenarios and realisations from a single model version, or using higher resolution to reduce some systematic biases. In a similar perspective, consideration may be given to sharing of computational resources between the different centres to optimise the design of multi-model ensembles.

2.7 Discussion and conclusions

Internationally coordinated projects of climate-change assessments have been increasingly common over the last decade or so. These projects consist in relatively large ensembles of simulations that use some population of models with a similar level of complexity in order to obtain climate-change projections according to different GHGA emission scenarios. While the real outcome for the future emission pathway is largely uncertain since mainly depending on the evolving socio-economical and political context, the divergence of projections obtained from several models also contributes importantly to our overall uncertainty about future climate changes (Hawkins and Sutton, 2009).

Beyond the sampling of a credible range of climate-change projections according to different emission scenarios, an important rationale that motivates the use of large ensembles such as the CMIP3 multi-model dataset is to obtain a collection of independent estimates of the future climate changes. The use of such a sample should result in two important benefits. First, a sample of independent estimates allows some cancellation of the errors across the different models and hence the ensemble average should converge toward the future climatic truth as the number of models contributing to the ensemble increases. The second benefit is that the spread between models' projections should be representative of the uncertainty about the climate projections.

It has been discussed throughout this chapter that there are several reasons to believe that climate models are not independent from one another. Moreover, very little is known on the extent to which models depend on each other since no measure of independence has been commonly accepted at this time. Several authors have assessed the independence of the projections by using the models' output (e.g. Jun et al. 2008b,a; Abramowitz and Gupta 2008; Knutti 2010; Pennell and Reichler 2011; Abramowitz 2010; Masson and Knutti 2011). Here we have adopted a different approach that aims at clarifying the concept of independence from the point of view of the models formulations. Of course, such an approach may become complicated since climate models have a very complex structure and include hundreds of parameters.

In order to explore the concept of the independence of the climate models *a priori* to their projections, we used as a starting point the assumption that models developed by a same institute share several characteristics at the structural, parameters and numerical levels. The structural level has been defined literally as the set of underlying physical assumptions that served as basis to each model. The way these assumptions are formulated, for example the choice of the parameterizations, has been also included in the structural level. Additionally, the values given to model's internal parameters and the numerical approximations have been highlighted as other types of model differences. It has been shown that the models (or versions) developed by an institute are prone to share such characteristics. Structural similarities are often straightforward to point out from models documentation provided by the PCMDI. Similarities in parameters and numerical characteristics are subtler and often not explicitly provided in the documentation. However, the fact that models from a same institute that differ only in a few components also suggests that the parameters and numerical approximations remain unchanged.

By paying attention to the consensus/disagreements in outputs as function of the degree of similarity in structural, parameters and numerical characteristics, we put in evidence that non-informative consensuses are likely to happen in a large multi-model ensembles

such as the CMIP3 multi-model dataset. The idea of non-informative consensuses has been explored in the specific context of the models developed by a same institute. However, it is very important to note that the non-informative consensuses are not limited to the same-institute context, but should rather be extended beyond this scope. Two models developed by different institutes with *a priori* no structural, parametric and numerical similarities could lead to non-informative consensuses since based on common physical assumptions and processes and interactions included in the models. For example, it is known (Knutti et al., 2010) that several CMIP3 models share common biases that are not exclusively limited to the same-institute context. On a larger scale, the generation of AOGCMs forming the CMIP3 multi-model dataset could also share important biases since none of them include a dynamical vegetation component, to cite but one example. A similar example is on the numerical assumptions for the models that use flux adjustment in the ocean (see Randall et al. 2007; Meehl et al. 2000). An important issue related to the independence of the climate models is that it depends on the simulated variable. For example, the two versions of the GISS model, EH and ER, share the same atmosphere, ice, land and coupling components, but differ in their ocean model components. Such a set-up is likely to result in a non-informative consensus in the surface air temperature over land, while informative disagreements (i.e. uncertainty) are found over the ocean. In this case, rather little information contributes to uncertainty over land, while an informative disagreement exists over the ocean.

The conceptual relationship between the prior (same-institute context) and posterior (consensus/disagreement in the outputs) definitions of independence does not appear sufficiently straightforward to be assumed blindly. In other words, one must find serious evidence to reject a consensus and hence to consider it as non-informative. Making such an inference should be done after paying attention to both the nature of the consensus (e.g. simulated variable, season, region and time period) and to the structural, parametric and numerical differences between the models (or versions). On the other hand, the direct application of this assumption by including only one model per institute in the scope of a specific study (e.g. Whetton et al. 2007) might be understood as little more

than an "educated guess" aiming at potentially decreasing the number of non-informative consensuses in the ensemble across the several variables, seasons, regions and time periods. This approach may be considered akin to a filtering or the assignment of zero weights to some of the models. This necessarily results in an increase of the standard error associated with the ensemble statistics since a smaller ensemble size is considered. A second potential benefit is reducing the risk of overconfidence in the ensemble statistics.

More technically, the rule of "one model per institute" may improve in some way the message conveyed by the ensemble statistics by reducing the risk of introducing artificial consistencies between models' projections. It is worth noting that further improvements are also possible since the issue of the models independence goes beyond of the same-institute context. Recent work has shown that the ensemble size of the CMIP3 multi-model dataset is much smaller than it appears from its number of participating models (e.g., Pennell and Reichler 2011, Annan and Hargreaves 2011). Moreover, given the relatively large sample of 24 models, simply removing the supplementary models for each centre might consist only in a slight reduction of the drawbacks related to the lack of independence between climate models. Using different methods for processing the models output, Pennell and Reichler (2011) estimated the effective number (N_{eff}) of climate models in the CMIP3 multi-model dataset to lie between 7.5 to 9, while Annan and Hargreaves (2011) obtained a range from 4 to 11. These estimates of the effective sample size can be combined into a single figure by their rounded average of 8 models. Now, by considering only the same-institute context as a proxy for model independence, the effective sample size of the CMIP3 multi-model dataset is estimated to 18 models when retaining only one model per institute (Tab. 2.1), with exception of the GISS family from which two models with different atmosphere components could be retained (AOM with EH or ER) and NCAR that provides two models (CCSM3 and PCM) with important differences in their response to identical climate forcing.

In order to understand the effect of these effective sample sizes on the statistics of the ensemble, let us recall the relationship of the standard error of the mean that can be

expressed as $\mathrm{Var}(\mu) = \sigma^2/N$, where μ and σ^2 are the true ensemble mean and inter-model variance and N the number of models in the ensemble. Assuming no changes in the mean and variance, comparing the perceived ensemble size (N) to the effective sample size (N_{eff}) corresponds to a standard error of the mean that is inflated by $\sqrt{N/N_{eff}}$. Hence, the ensemble mean is inflated by 73% for $N_{eff} = 8$ while by only 15% when using $N_{eff} = 18$. With similar arguments, this reasoning may as well be applied to the standard error of the inter-model spread that is sometimes used to assess the model uncertainty. From the point of view of mitigation and adaptation strategies to climate change, overestimating the uncertainty of projections or the standard error of the ensemble statistics is generally more cautious, but indeed more expensive, since a broader range of climate outcomes are considered. On the other hand, underestimating these ranges certainly simplifies the mitigation process while increasing the risk of an outcome that lies outside the measured range of uncertainty, and hence that might be unexpected by the mitigation plan.

In the climate modelling community, the democratic way of thinking the message conveyed by an ensemble such as "one model, one vote" has been discussed and questioned in the context of model performance. Since models perform differently in reproducing the various facets of the climate system, it is sometime argued that climate projections should be weighted according to some performance criteria. While model democracy would be somewhat compromised by using such an approach (Knutti, 2010), the same concept could be extended according to the issue of model independence as an "institutional democracy" that should already exist in the ensemble results or being imposed through the analysis of the results. A potential way to induce institutional democracy in the ensemble data would be to invite the participating centres to "demonstrate" the benefits of including a second model version as how it would contribute by potentially adding some value to the ensemble. The latter way corresponds to a post-filtering or weighting of the ensemble's models according to some independence metric, which consists in an emerging field of research. Since the issue of model independence goes far beyond the same-institute context, one could argue ultimately that the climate modelling science

should tend toward some kind of "ideological democracy" since even different research institutes may not be sufficiently independent from each other. The latter considerations could help spanning a broader and more realistic range for the uncertainty around the climate-change modelling problem, whether according to pre-selected participating institutes or to post-filtering an ensemble of opportunity.

Of course, the climate modelling science is not the only field of research where independence matters. In its precursory works, Levins (1966) noted that the use of several different biological population models may lead to a same result despite their different underlying assumptions. The author used the terminology of a "robust theorem" for a result that is free of the details of each model. This concept can be interpreted similarly to that of an informative consensus discussed in this chapter.

Appendix 2.A : Statistical significance of the difference between two ensemble means (t-test)

Let μ_X and μ_Y be the ensemble mean climate-change signals for two models denoted by the X and Y indices. The latter consist in true means that could be estimated by using a sufficiently large number of realisations differing in the initial conditions. By simplicity, we firstly assume equal inter-member variances between the models, i.e. that $\sigma_X = \sigma_Y$. The null hypothesis of equal means, H_0, can be defined as

$$H_0 : \mu_X = \mu_Y \tag{2.1}$$

and the t statistics

$$t = \frac{\hat{\mu}_X - \hat{\mu}_Y}{\hat{\sigma}_p \sqrt{\frac{1}{N_X} + \frac{1}{N_Y}}} \tag{2.2}$$

where N_X and N_Y correspond to the models' sample sizes and $\hat{\sigma}_p^2$ to the pooled variance

$$\hat{\sigma}_p^2 = \frac{\sum_{i=1}^{N_x}(x_i - \hat{\mu}_X)^2 + \sum_{i=1}^{N_y}(y_i - \hat{\mu}_Y)^2}{N_X + N_Y - 2} \tag{2.3}$$

with $N_X + N_Y - 2$ degrees of freedom.

In a case where the variances can not be assumed as equal ($\sigma_X \neq \sigma_Y$), the t statistics becomes

$$t = \frac{\hat{\mu}_X - \hat{\mu}_Y}{\sqrt{\frac{\hat{\sigma}_X^2}{N_X} + \frac{\hat{\sigma}_Y^2}{N_Y}}}, \tag{2.4}$$

that consists in an approximation of the t distribution with its number of degrees of freedom being estimated from the data such as

$$df = \frac{\left(\hat{\sigma}_X^2/N_X + \hat{\sigma}_Y^2/N_Y\right)^2}{\frac{\left(\hat{\sigma}_X^2/N_X\right)^2}{N_X - 1} + \frac{\left(\hat{\sigma}_Y^2/N_Y\right)^2}{N_Y - 1}}. \tag{2.5}$$

The critical values for a given significance level can hence be found using a table of the t

distribution. More details on these tests and the related tables can be found in common statistical textbooks such as von Storch and Zwiers (1999) and Wilks (2011).

Tab. 2.1: Name of the research institutes/groups that provided several models or versions to the CMIP3 multi-model archive.

Name of the research institute/group	Country	Acronym
Canadian Centre for Climate Modelling & Analysis	Canada	CGCM
Center for Climate System Research (The University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC)	Japan	MIROC
The Commonwealth Scientific and Industrial Research Organisation (Atmospheric Research)	Australia	CSIRO
US Dept. of Commerce / NOAA / Geophysical Fluid Dynamics Laboratory	USA	GFDL
NASA / Goddard Institute for Space Studies	USA	GISS
National Center for Atmospheric Research	USA	NCAR
Hadley Centre for Climate Prediction and Research / Met Office	UK	UKMO

Tab. 2.2: Table of the main structural, parameters and numerical differences between pairs of models developed by a same research institute within the CMIP3 multi-model archive. Models are compared according to their main components : atmosphere (A), ocean (O), sea ice (I), coupling (C) and land surface (L). The differences are categorised as resolution (R), version (V), model (M) and no change (-).

Pair	Acronym	Models	Difference	A	O	I	C	L
I	CGCM	T47 vs T63	Change in Δ_{xy} for At. and Oc.	R	R	-	-	-
II	MIROC	T106 vs T42	Change in Δ_{xy} for At. and Oc.	R	R	-	-	-
III	CSIRO	3.0 vs 3.5	Oc. eddy parameterization (transport coefficient) & mixed-layer treatment (turbulent kinetic energy), sea ice (numerical scheme), coupling (wind stress), treatment of surface runoff and river routing scheme	-	V	V	V	V
IV	GFDL	CM2.0 vs CM2.1	Numerical scheme: advection, gravity waves and damping at the top boundary for At. and leapfrog timestepping vs staggered for Oc.	V	V	-	-	-
V	GISS	EH vs ER	Different Oc. (HYCOM vs Russell)	-	MR	-	-	-
VI	GISS	AOM vs ER	Different At., sea ice, coupling and land, different versions of the Russell Oc.	MR	VR	M	M	M
VII	GISS	AOM vs EH	Different At., sea ice, coupling and land, different Oc. (Russell2 vs HYCOM)	MR	MR	M	M	M
VIII	NCAR	CCSM3 vs PCM	Different models developped by the same institute (NCAR).	MR	MR	M	M	M
IX	UKMO	CM3 vs GEM1	Different models developped by the same institute (Hadley Centre).	MR	MR	M	M	M

Tab. 2.3: Feasibility of a t-test for the difference between ensemble means of different pairs of models, according to the number of simulations available for the A1B scenario within the CMIP3 multi-model dataset. Sample sizes of the two models in a pair are denoted by N_X and N_Y. In the last column (t), the pairs are denoted by "0" when the test can not be performed, by "E" when equal variances have to be assumed and by "U" when unequal variances can be considered.

Pair	Name	N_x	N_y	t
I	CGCM (T47-T63)	5	1	E
II	MIROC (T106-T42)	1	3	E
III	CSIRO (3.0-3.5)	1	1	0
IV	GFDL (2.0-2.1)	1	1	0
V	GISS (EH-ER)	3	4	U
VI	GISS (AOM-ER)	2	4	U
VII	GISS (AOM-EH)	2	3	U
VIII	NCAR (CCSM3-PCM)	7	3	U
IX	UKMO (CM3-GEM1)	1	1	0

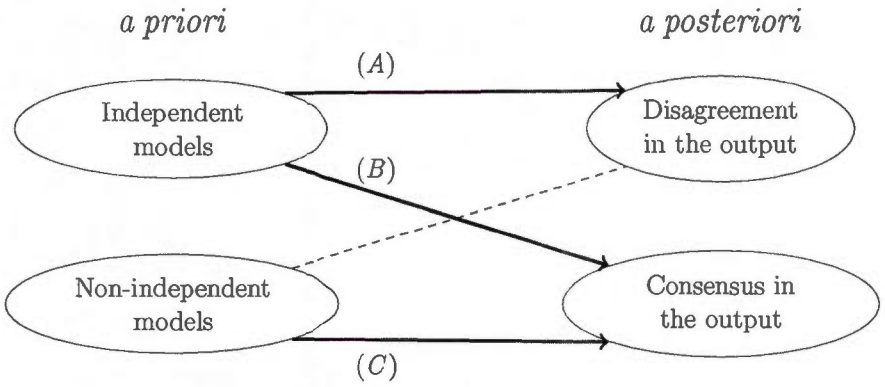


Fig. 2.1: Schematic of the conceptual relationship between prior and posterior definitions of model independence.

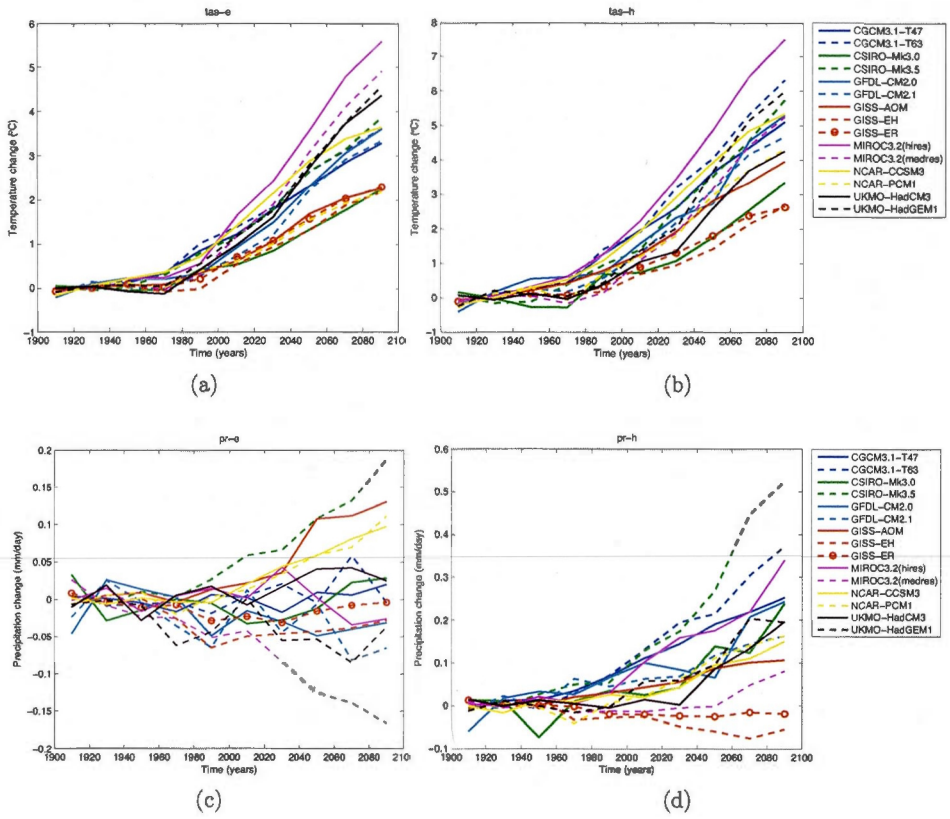
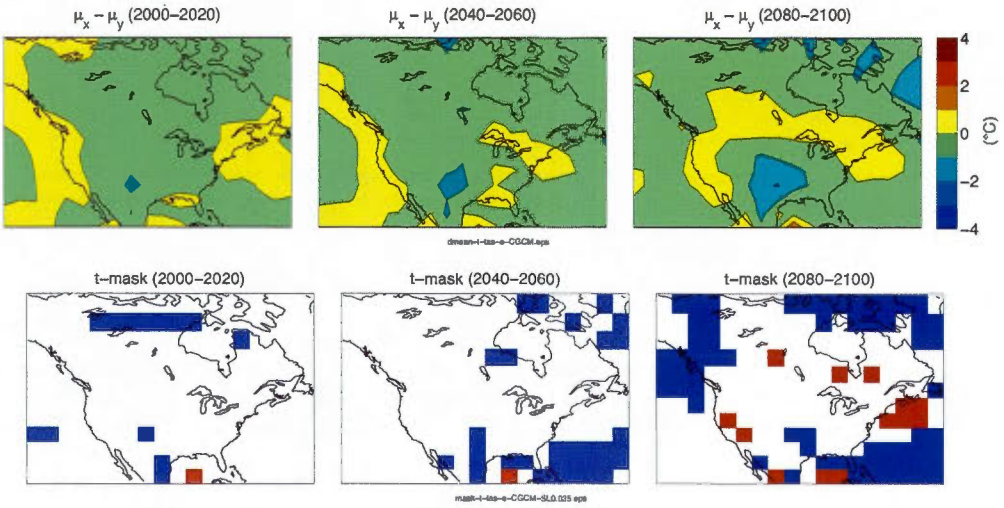
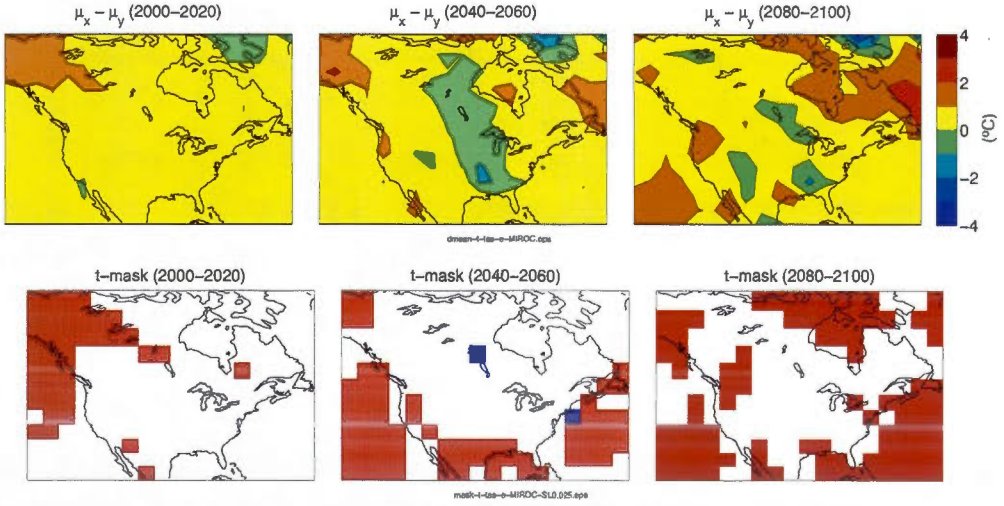


Fig. 2.2: Climate-change projections for the a) summer and b) winter surface air temperature and for the c) summer and d) winter precipitation rate. These changes are calculated over 20-year time periods compared to the 1900-1950 level for each of the models presented in Tab. 2.2. All available realisations are averaged over the regional domain of North America.

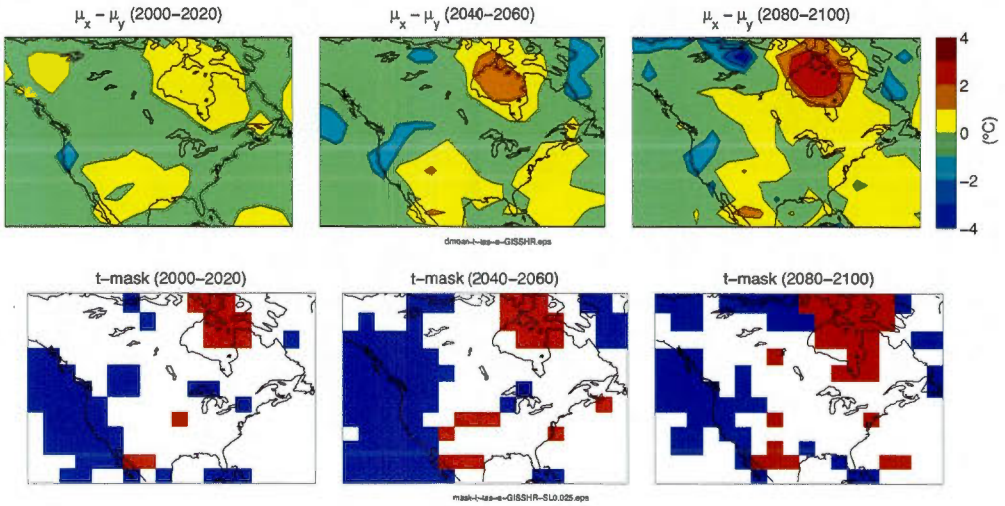


(a) CGCM : T47-T63

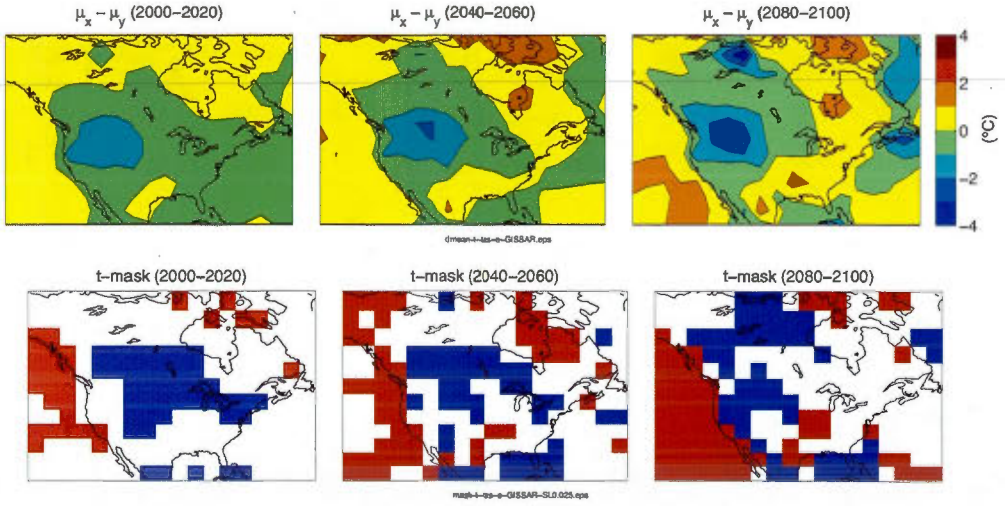


(b) MIROC : T106-T42

Fig. 2.3: To be continued...

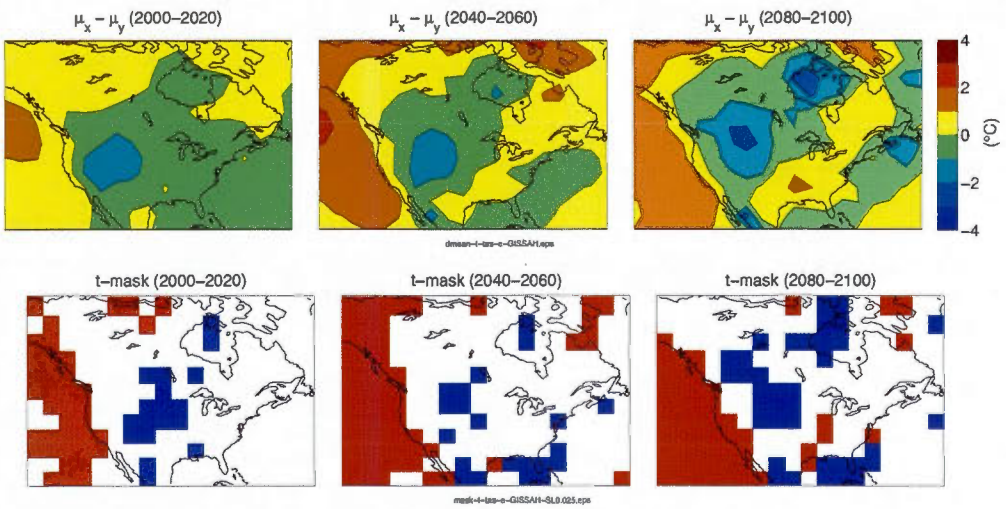


(c) GISS : EH-ER

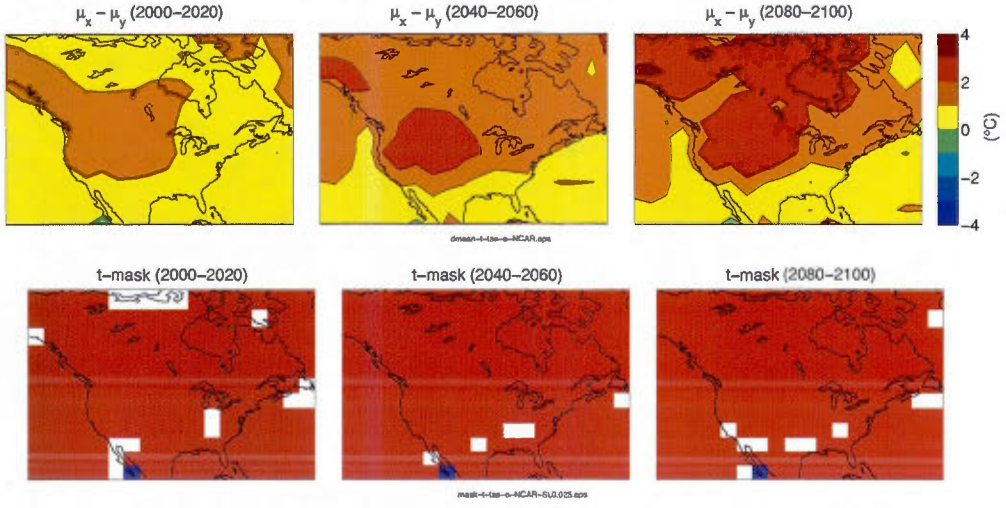


(d) GISS : AOM-ER

Fig. 2.3: To be continued...



(e) GISS : AOM-EH



(f) NCAR : CCSM3-PCM

Fig. 2.3: Difference of the ensemble mean climate-change signal for different pairs of models (or versions) developed by the same research institute. The climate-change signal is calculated for each simulation relatively to the 1900-1950 period. The panel at the bottom of each difference shows the mask of rejection of the null hypothesis by using a two-tailed t-test at the 5% significance level (2.5% on each side). Red and blue colours mean positive and negative differences respectively.

CHAPTER III

THEORETICAL FRAMEWORK FOR RECONSTRUCTING MISSING MEMBERS IN A MULTI-MODEL ENSEMBLE OF AOGCMS

ABSTRACT

Model Intercomparison Projects aim to compare climate-change projections obtained from different modelling centres. The main value of such ensembles of simulations resides in providing the scientific community with the plausible range of future climates. However, such ensembles are often constructed in a rather arbitrary manner, mainly based on the computing resources available to the participating centres. It follows that studying the uncertainty in such ensembles can suffer of limitations due to the use of a non-systematic experimental framework. In order to circumvent these limitations, one can consider the alternative of artificially regenerating the "missing simulations" in the ensemble in order to provide a systematic framework for the further analysis. The present chapter investigates the feasibility of two data-reconstruction methods : the single-model and multi-model pooling. The first method consists in regenerating new members for a model by using only the information available from that model. The second method consists in regenerating members by using the information available from several models in the ensemble. The choice of the method depends on 1) the invariance of the statistics when calculated over time or across the multiple realisations associated to a model (ergodicity), and 2) the similarities in inter-member (internal) variability across models.

3.1 Introduction

Over the last years, several Model Intercomparison Projects (MIPs) have been conducted internationally by the climate modelling community in order to characterise the main sources of uncertainty affecting the climate-change projections for the 21st century. Among these sources, the uncertainty emerging from different modelling approaches

(commonly known as "model uncertainty") can be investigated through the use of multi-model ensembles (Tebaldi and Knutti, 2007). One basic characteristic of such ensembles is that different models are run under similar external forcings; popular examples are the Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset (Meehl et al., 2007a,b) and CMIP5. Commonly applied external forcings include anthropogenic sources such as the greenhouse gases and aerosols (GHGA) emissions and land-use scenarios (Nakicenovic et al., 2000). On the other hand, forcings from natural sources are also applied to climate models simulations as the events of volcanic emissions (Sato et al., 1993; Ammann et al., 2003) and the historical trends and cycles in the solar irradiance (Lean et al., 1995). The external forcings that are prescribed in the models are also affected by uncertainty; for example, the anthropogenic forcings are intimately related to the future socio-economical and political context.

Another source of uncertainty is the natural variability of the climate system that ranges over broad time scales, from seconds to thousands of years. It is generally considered as a source of uncertainty that is internal to the system since it appears even under stationary climate forcing. The natural variability can be sampled by using a single but long climate-model simulation (e.g. thousands of years). Similarly, the inter-member variability appears as the spread between several realisations using the same model, but with slight differences in the initial conditions. This measure of spread is generally attributed to the natural climate variability when sampled from an Atmosphere-Ocean General Circulation Model (AOGCM) (Sorteberg and Kvamstø, 2006; Deser et al., 2010). Comparatively, the inter-member variability that is sampled from a Regional Climate Model (RCM) (Alexandru et al., 2007; Lucas-Picher et al., 2008; Nikiema and Laprise, 2011) is generally smaller in magnitude. While the realisations from an RCM use the same boundary conditions that also include the natural climate variability (e.g. from an AOGCM), the inter-member variability represents in this case the deviations from an externally forced state.

For a given external forcing scenario, a multi-model ensemble (MME) implies simulations

from several models and where each one is represented by one or several realisations (members). Different number of members is often noted across models due to the high cost of producing simulations for long periods (e.g. hundreds of years) with increasingly high resolutions, while the participating centres have limited resources and their own interests. Such a MME can be represented as a two-dimensional matrix of simulations (models, members for each model) where some elements are "missing" compared to an idealised ensemble where each model would have the same number of members.

Even more complex MMEs result from dynamical downscaling with Regional Climate Models (RCMs) driven by lateral atmospheric and sea-surface boundary conditions from AOGCM simulations (e.g. ARCMIP [1], NARCCAP [2], ENSEMBLES [3], CORDEX [4]). Such framework also suffers from missing matrix elements due to the very high cost involved in attempting to downscale each member of each AOGCM with each RCM. However, the effects of the missing model combinations can be minimised when the missing elements are systematically distributed across the matrix. For example in the NARCCAP project, each AOGCM is used to drive three RCMs, while each RCM uses boundary conditions from two AOGCMs (Mearns et al., 2009). The ENSEMBLES project on the other hand, while attempting to account for several sources of uncertainty, suffered from an imbalance between the sampling of the scenarios, model combinations and member realisations (van der Linden and Mitchell, 2009).

The use of an unbalanced ensemble can involve biases and large sampling errors when partitioning the uncertainty into several components of variability. In order to circumvent these issues and hence reinforce the message conveyed by the analysis of the uncertainty

---

1. The Arctic Regional Climate Model Intercomparison (ARCMIP), `http://curry.eas.gatech.edu/ARCMIP/`.

2. North American Regional Climate Change Assessment Program (NARCCAP), `http://www.narccap.ucar.edu/`.

3. The ENSEMBLES Project, `http://www.ensembles-eu.org/`.

4. COordinated Regional climate Downscaling Experiment (CORDEX), `http://www.meteo.unican.es/en/projects/CORDEX/`.

components, Déqué et al. (2007, 2012) used data-reconstruction methods for projecting a non-systematic framework onto a systematic one. One should keep in mind though that an inappropriate data-reconstruction method could also result in an increase of the uncertainty by adding arbitrary noise to the original dataset.

Since several empirical ways can be imagined in order to regenerate the missing elements of a matrix of simulations, this chapter presents a theoretical framework that aims at choosing the most suitable between two data-reconstruction methods. The first method, henceforth referred to as single-model pooling (SMP), consists in a resampling from the available realisations of one model in order to generate artificial members to that model. The second method, referred to as multi-model pooling (MMP), uses the realisations from different models to generate artificial members to any of the models. The choice of the most suitable approach implies two scientific questions that are inherent to the nature of the MME. The first one (Q1) asks for ergodicity in a single-model ensemble, i.e. the invariance of the statistics whether computed over time or members (Peixoto and Oort, 1992). The second (Q2) addresses the differences in the simulated climate variability by different models, more precisely whether these differences are physically significant.

Sect. 3.2 presents the theoretical framework in order to make an educated guess on which data-reconstruction method is the most appropriate for generating the missing members in a MME. Sect. 3.3.1 describes the data used in this study, which consist in an subset extracted from the CMIP3 multi-model dataset (MMD). Sect. 3.3.2 presents the decomposition of the climate variability into forced and unforced components, and Sect. 3.3.3 develops the testing frameworks related to both questions Q1 and Q2. The results are analysed in Sect. 3.4, followed by a general discussion in Sect. 3.5.

3.2 General approach to member reconstruction

As noted in Introduction, the use of a non-systematic ensemble framework can be an issue when applying common statistical methods (e.g. analysis of variance) to separate

the different components of the uncertainty. In this chapter, we propose two methods for generating the missing elements in an unbalanced ensemble such as the MMD. Basically, both methods generate artificial simulations by resampling over a "pool of climatic data", either from single-model or multi-model information. We will first describe the two types of pool, and then we will integrate the two approaches within a single theoretical framework aiming to make an educated guess on which one is the most suitable for applying to the MMD.

An important part of our analysis is based on the fact that a single-model ensemble of climate simulations can be considered as *ergodic*, i.e. that the main statistics (e.g. mean and variance) are invariant whether calculated over time for one realisation or over several members for one specific time. In order to apply the ergodic assumption to the case of a single-model ensemble of AOGCM simulations, some basic conditions have to be met. Firstly, the realisations have to be run under stationary conditions or more specifically, with constant external forcings applied (e.g. GHGA emission and land-use scenarios, volcanic emissions, solar irradiance). Secondly, the realisations are independent, that is the initial conditions have been forgotten by leaving a sufficiently long spin-up time period at the beginning of each simulation. Finally, the ensemble size and the length of the simulations are sufficiently large. It is worth noting that seasonal, annual or longer time-averages can be considered in order to avoid any correlation between the members due to either the daily and annual cycles. For more details about the previous conditions, Appendix 3.A presents an example of application of the ergodic assumption to a single-model ensemble of AOGCM simulations.

By assuming the truthfulness of the ergodic assumption for a single-model ensemble of simulations, it follows that artificial simulations can be regenerated for one model by random sampling over the data available from the members of that model. By using such an approach, the artificial time series are not expected to reproduce all the characteristics of a climate model simulation, such as the sequence of weather events. However, under the ergodic assumption, the statistics are expected to be preserved in the extended

ensemble : i.e. the mean or variance of climate calculated for one realisation should equal the ensemble mean at any specific time. In what follows, we refer to this method as the single-model pooling (SMP) since the information available from one model is used as a pool for generating artificial members for that model.

The SMP method can be extended by pooling together the realisations from several models in order to form a multi-model pool (MMP) from which climatic data can be resampled. While the SMP method could be limited in the case of a model providing very few members (e.g. one), the MMP method offers a wider pool to pick from. Unfortunately, it is well known that different models can show relatively large differences in their simulated climate (Greene et al. 2006, Gleckler et al. 2008). Inter-model differences can also be expected for the natural variability, but the physical significance of these differences could be judged small enough in some cases to consider the MMP approach. Assuming the model biases to be removed and thus imposing equal means between the models, the MMP generates artificial members for one model by sampling over the climatic data available from the members of other models.

As stated previously, the two pooling methods involve an ensemble of simulations run under stationary condition. Obviously, this condition is not met for the MMD projections into the 21st century where the models employ transient external forcings as GHGA emission and land-use scenarios, volcanic emissions, variations in the solar irradiance, etc. However, one can approach stationary conditions (and hence ergodicity) by removing the forced component in the simulations. Such an approach allows leaving only the unforced component that represents the internal variability as simulated by the models. It is worth noting that the forced component can be extracted by using the ensemble mean, but doing so generally necessitates a large number of realisations (e.g. Wigley et al. 2005). In the case of the MMD that generally provides very few members for each AOGCM, an important part of the forced component can be removed by detrending the simulations according to the ensemble mean of each single-model ensemble (see Appendix 3.B).

Fig. 3.1 presents a flowchart that summarises the proposed theoretical framework for

reconstructing the missing members in a MME of AOGCMs in order to obtain a balanced ensemble framework. Beginning from the top of the diagram with an MME under transient forcing, the simulations have to be detrended to approach stationary conditions and hence allows the ergodic assumption to apply. The next step is represented in the figure by a diamond box that involves a first scientific question that is denoted as Q1 ; this question aims at verifying if the detrending of the simulations satisfies the ergodic assumption. It involves a test (Appendix 3.C) that allows detecting if a single-model ensemble can be treated as ergodic by investigating both the statistical and physical significance of the non-ergodic part of the signal. In the case that the single-model ensemble is judged ergodic, the next step in the flowchart involves a second scientific question, denoted as Q2, which checks the equality of the climate variability as simulated by different models. To help answering this question, a second test (Appendix 3.D) is constructed for evaluating both the statistical and physical significance of the inter-model differences in the simulated climate variability.

In the case that the answer to both questions Q1 and Q2 is "yes", we consider as suitable the use of a MMP for reconstructing the missing elements in the MME. In the case that we answer "no" to Q2, a SMP should be preferred. Since ergodicity is expected for a single-model ensemble run under stationary conditions (Appendix 3.A), it is worth noting that answering "no" to Q1 could reflect an inappropriate detrending of the simulations. For example, it could be due to the degree of the fitted polynomial function that is not appropriate or that a correction should also be applied to higher statistical moments. An alternative could be to reject the non-ergodic simulations or to consider other methods for data reconstruction. Once the choice of the most suitable type of pooling is done, artificial members can be generated for any model in the ensemble. It is worth noting that the removed trends can be "added back" to the generated members, depending on the needs of further analysis.

3.3 Experimental framework

3.3.1 Data

In this study, we consider a subset of simulations from the CMIP3 MMD. More specifically the eleven models providing more than a single realisation for the A1B scenario have been retained. This subset from the large ensemble will be referred to as the MME. The projections for the A1B scenario have been merged with their corresponding run for the 20th century, resulting in 42 simulations covering the period from 1900 to 2100. The model names are shown in Tab. 3.1 with their respective number of available members (N_K). Also are shown in the table which models include the radiative forcing due to the volcanic emissions and variations in the solar irradiance. The complete models' specifications can be found on the Program for Climate Model Diagnosis and Intercomparison (PCMDI) website at `http://www-pcmdi.llnl.gov`. The present study focuses on the time evolution of the summer-average surface air temperature over North America. The simulations from the different models have been linearly interpolated over a common grid of $4° \times 5°$ degrees. The time series are detrended according to 4th-degree polynomial functions fitted to the ensemble mean of each model. This method for detrending the simulations is detailed in Appendix 3.B.

3.3.2 Components of variance

Fig. 3.2 schematises an ensemble of simulations performed by a single climate model. This single-model ensemble can be seen as a matrix (X) containing time periods (t) and realisation (member) number (k). Note that the model index (m) and spatial coordinates are implicitly considered to lighten the notation. An element (X_{tk}) can be described as

$$X_{tk} = \mu + a_t + e_{tk}, \tag{3.1}$$

where μ represents the mean climate of the single-model ensemble. As described in Appendix 3.B, the time series are detrended in order to approach stationary conditions.

It follows that the X_{tk} are time deviations (from the trend) and hence that μ tends to zero by construction. The first component (a_t) describes the deviations that are shared across all the realisations according to the time. Because of the detrending, this component does not reflect the GHGA or aerosols emissions, but includes faster cycles resulting from volcanic emissions and solar irradiance when these are taken into account and simulated by a model.

In the following analysis, we assume a_t as a random effect occurring in time with some variability defined as the forced variance σ_F^2. For e_{tk}, it consists in the residual fluctuations that are assumed to be independent and identically distributed (iid) according to t and k. The residual fluctuations component has a variance σ_{IV}^2 and is expected to represent the internal variability as simulated by the model. Based on the assumption that a_t represents the forced component, σ_{IV}^2 can be also interpreted as the natural variability of the modelized climate system under stationary conditions. Based on the previous statistical model and its related assumptions, the elements X_{tk} are distributed with a variance that will be referred to as the total climate variability (σ_{tot}^2) :

$$\sigma_{tot}^2 = \sigma_F^2 + \sigma_{IV}^2, \tag{3.2}$$

a sum of the forced variance (σ_F^2) and the internal (natural) variability (σ_{IV}^2).

3.3.3 Hypotheses testing

The two questions (Q1 and Q2) that appear in the flowchart in Fig. 3.1 can be investigated through the use of test statistics. Each of these statistics involves the rejection of a null hypothesis that translates the scientific question under study.

A formulation of the test associated to the first question (Q1) can be found in Appendix 3.C. The null hypothesis is denoted as H_0^{ergo} and states that there is no forced component of variability according to the time and hence that all the variability of the single-model ensemble is described by the unforced component (σ_{IV}^2). A rejection of H_0^{ergo} (using an

F-ratio denoted as F_1) means that the forced component of variability is statistically significant, i.e. that the single-model ensemble is not ergodic from the point of view of the test. It follows that the ergodic assumption cannot be verified directly but can only be rejected at some significance level. In order to appreciate the physical significance of the regions where the null hypothesis is rejected, P_1 represents the ratio between the forced variability (σ_F^2) and the total climate variability ($\sigma_F^2 + \sigma_{IV}^2$).

The second question (Q2) is addressed in Appendix 3.D and can be translated by the null hypothesis H_0^{var}. This hypothesis states that the simulated total climate variability (σ_{tot}^2) is equal between two models (labeled as m and m'). In order to reject this hypothesis, we use the statistics $F_2(m, m')$ that consists in an F-ratio between the two variances. In order to quantify the physical significance of the difference in variability, the relative error of variance $P_2(m, m')$ is used, which consists in a ratio of the difference in variance with the mean variance of the two models (Eq. 3.16 in Appendix 3.D).

3.4 Results

In the following two sections, we evaluate the feasibility of applying the SMP and MMP member-reconstruction methods in order to regenerate the missing simulations in the MME. The results are presented through an investigation of the two scientific questions (Q1 and Q2) that are involved within the theoretical framework presented in Fig. 3.1.

3.4.1 Ergodicity in single-model ensembles

In this section, we present the results related to the first question (Q1) of the theoretical framework proposed in Fig. 3.1. This question focuses on the ergodic assumption applied to single-model ensembles. In our analysis, we use the summer mean surface air temperature for the eleven models of the MMD that provide more than a single realisation (Tab. 3.1). The results for this test are shown for the 20th and 21st centuries separately.

Considering first the simulations over the 20th century, in Fig. 3.3 is shown the variance

ratio (P_1) for each of the models that are sorted from the largest (Fig. 3.3a) to the smallest (Fig. 3.3k) number of realisations. The coloured regions indicate the areas where H_0^{ergo} is rejected with a significance level of 10%. Given the significance level, the critical values of F_1 are calculated from the F-distribution and are determined by both the number of climatic time periods N_T and ensemble size N_K as $F(N_T - 1; N_T \times (N_K - 1))$ (see Appendix 3.C). For example, the critical value of the variance ratio P_1 is 0.03 for the CCSM3 model (Fig. 3.3a) and increases to 0.13 for GISS-AOM (Fig. 3.3k) due to the reduction in the ensemble size from $N_K = 7$ to $N_K = 2$. It is worth noting that a lower significance level (e.g. 1%) would have a similar effect by increasing the prominence of the non-rejecting regions (in white).

The rejecting rate can be defined as the percentage of the domain where the null hypothesis is rejected. By comparing the different models in Fig. 3.3, it can be seen that some models show rejection rates that are higher than 20% of the domain (CCSM3, GISS-ER, PCM, GISS-EH and MIROC3.2(medres) in Fig. 3.3a, d, f, g and i respectively). The other models show relatively small rejection rates with values smaller than 9% of the domain.

If we now pay attention to the spatial distribution of the ratio (P_1) between the forced and total variability (referred to as the non-ergodic signal), in Fig. 3.3a, f and g, there are areas where $P_1 > 20\%$ of variance. This mainly occurs in the southeast part of the domain, over the Gulf of Mexico and extending over the Atlantic Ocean. Weaker signal is also noted along the east-coast of the United States for MIROC3.2(medres) and ECHO-G (Fig. 3.3i and j), and over the Gulf of Mexico for GISS-ER (Fig. 3.3d). An interesting feature is that the detection of a non-ergodic signal generally appears over the oceanic or coastal regions while less occurrence appears over the land regions where the non-ergodic signal is generally smaller than 10% of the variance, with an exception for the PCM and GISS-EH models showing a local maximum ($P_1 \approx 15\%$) over the Québec province of Canada. Recalling Tab. 3.1, the models without volcanic and solar forcing are CGCM3.1(T47), MPI-ECHAM5, FGOALS-g1.0, and GISS-AOM. In Fig. 3.3 (b, e, h

and k), all four models do not display any clear signal of rejection of the null hypothesis (H_0^{ergo}) through very low rejecting rates. It is worth noting that the MRI-CGCM2.3.2 model does include both of the volcanic and solar forcing agents but does not show any significant variability according to the present test.

The results for Q1 are also shown for the 21st century in Fig. 3.4. The occurrence of a non-ergodic signal is relatively rare for most of the models and the rejection rate is always smaller than 12% of the domain. Also, the non-ergodic signal present near the Gulf of Mexico in the simulations for the 20th century does not appear at all in the 21st century. This result shows that after detrending the simulations, some variability due to external forcings is remaining in the simulations for the 20th century but not in those for the 21st. It seems that this remaining variability is mainly due to the volcanic emissions events in the 20th century rather than to the cycles in the solar irradiance (Appendix 3.B). However, the effect of these two forcings is difficult to evaluate separately by using the MMD since the models that include volcanic forcing also account for variations in solar irradiance.

Recalling that the critical value of the F-distribution depends on the number of climatic time periods (N_T) and the ensemble size (N_K), it is worth noting that a large N_K does not necessarily involve a high rejecting rate of the null hypothesis. For example, the CGCM3.1(T47) and MRI-CGCM2.3.2 (Fig. 3.3b and c) models that have relatively large ensemble sizes ($N_K = 5$), H_0^{ergo} is rejected over only 5% and 2% of the domain respectively. Given the level of statistical significance, a single-model ensemble can be considered as non-ergodic from a statistical point of view over the regions where the null hypothesis H_0^{ergo} is rejected. From the physical point of view, one can argue for ergodicity over the same regions if the variance ratio (P_1) is judged sufficiently small. Overall, the results suggest that we can assign a rather positive answer to Q1, i.e. that the ergodic assumption generally holds over a great part of the domain and especially for the land regions in the simulations for the 20th century and practically for the entire domain in the 21st century.

3.4.2 Inter-model differences in the simulated total climate variability

In the previous section, we analysed the non-ergodic part of the signal by comparing the forced variability (σ_F^2) to the total climate variability ($\sigma_F^2 + \sigma_{IV}^2$). The forced component generally appeared relatively small and hence the internal variability (σ_{IV}^2) has to count for the largest part of σ_{tot}^2 due to the sum of variances (3.2). In this section, we investigate both the statistical and physical significance of the inter-model differences in the simulated total climate variability. Such an analysis can provide valuable information for answering question Q2, which plays a decisive role in the selection of a data-reconstruction method according to Fig. 3.1. As shown in Appendix 3.D, the total climate variability simulated by two models, denoted as $\sigma_{tot}^2(m)$ and $\sigma_{tot}^2(m')$, can be compared using the relative error of variance $P_2(m, m')$. Using the 11 models of the MME, 55 subsets of two models can be formed. These 55 comparisons are presented in Fig. 3.6 in the form of a strictly upper triangular matrix of panels where the m models are represented as rows (a to j) and the m' as columns (b' to k') (see Tab. 3.1 for the model name associated to each letter). The P_2 statistics is bounded between $P_2 = -2$ (in blue) where the ratio $F_2 = 0$ and saturates to $P_2 = 2$ (in red) for $F_2 \to \infty$. For equal variability between the two models (i.e. $F_2 = 1$), the relative error is $P_2 = 0$. In Fig. 3.5, P_2 is plotted as function of F_2 according to Eq. 3.16 in Appendix 3.D. Using a two-tailed F-test at the 10% significance level, a white mask has been applied over regions where the difference of variances is not statistically significant. In these regions, not enough evidence allows to distinguish the two models' climate variability.

In Fig. 3.6, the inter-model comparison of the total climate variability is done for the 20th century. A general feature is that the rejection rate of the null hypothesis (H_0^{var}) is rather large, i.e. that the inter-model differences in the total climate variability are statistically different over large proportions of the domain. It is worth noting that the sign of $P_2(m, m')$ is determined by the order of the comparison between m and m'. For example, if we focus on the comparison of GISS-EH with the other models, we follow the g' column to be compared with the models a to f and continue on the row g for a

comparison with the models h' to k'. For this model, one can identify a maximum value of relative error located over the Hudson Bay represented in blue in the g' column and in red in the g row. Particularly, the panel (g, h') displays a maximum corresponding to a total climate variability that is 10 times larger in variance for GISS-EH than for the FGOALS-g1.0 model. Inversely the GISS-AOM model has generally smaller total climate variability then the other models over the largest part of the domain where the values of P_2 are positive (column k'). An interesting feature can be seen in the pair (d, g') that shows a relatively low rejection rate compared to the other pairs. These models (GISS-ER and GISS-EH) are developed by the same institute but differ only in their ocean component. This could probably explain the maximum of relative error of variance ($P_2 \approx -1.5$) found over the Hudson Bay. On the other hand, the continental values generally do not exceed $P_2 = 0.5$, which corresponds approximately to twice the variance of the reference (i.e. $F_2 = 2$ according to Fig. 3.5). These results are consistant with the inter-model comparison done by (Santer et al., 2011).

If we now look at the cross-model comparison for the 21st century (Fig. 3.7), the results appear very similar to those obtained for the 20th century (Fig. 3.6). For example, the positive maximum value of relative error over the Hudson Bay is preserved for the GISS-EH model (g) as for the GISS-AOM model (k) that has a total climate variability that is smaller than the other models over a great part of the domain. Also, the total climate variability of the GISS-ER and GISS-EH models are still statistically similar over the land since the null hypothesis is weakly rejected.

As noted in Sect. 3.4.1, the non-ergodic part of the variability (σ_F^2) can greatly contribute to the total climate variability (σ_{tot}^2). However, the large contributions are generally located in an area characterised by rather small total climate variability such as the Gulf of Mexico (not shown). Added to the fact that the inter-model ratios of total climate variability do not change much between the 20th and 21st centuries, this suggests that the internal variability of the climate models is relatively robust over centennial periods under the A1B emissions scenario. However, investigating shorter periods would probably

reveal temporal changes in the simulated internal variability as obtained by Räisänen (2002) using the previous generation of models.

3.5 Discussion and conclusions

Model intercomparison projects (MIPs) consist in internationally conducted experiments where different modelling centres provide simulations from one or several models. In order to encourage diversity in the participating models, low requirements are generally asked to the centres in the number of simulations to be provided. It follows that these ensembles are very likely to result in non-systematic frameworks due to an imbalance in the sampling of scenarios, models and realisations. An unbalanced ensemble design can induce errors in the use of some diagnosis tools (e.g. analysis of variance). We proposed two simple methods in order to artificially generate the members that are missing in a multi-model ensemble in order to obtain a balanced framework. Both methods use a pool of climatic data which are resampled to create artificial time series. The first method involves a pool constructed using the realisations from a single-model, the second use data from multiple models.

The single-model pooling (SMP) method requires that the single-model ensemble is ergodic and hence that time periods can be considered in the construction of new members and vice versa. The CMIP3 multi-model dataset (MMD) being run under transient forcings, the most important being the GHGA emissions, the simulations have been detrended in order to remove the main part of the forced component. It appeared from the results that the single-model ensembles are rather ergodic even if the effect from some transient forcings has survived to the detrending. Especially, the simulations for the 21st century appeared more ergodic since subject to less synchronised transient forcings compared to the 20th century where the volcanic emissions and the solar irradiance are modulated in time based on historical records. The non-ergodic signal for the 20th century has been mainly detected over the Gulf of Mexico for several models while this feature does not appear for the 21th century.

In order to use the multi-model pooling (MMP) method, the ergodic assumption must apply for the different single-model ensembles and with the supplementary condition that the climate variability must be simulated with a similar intensity among the models. The results show that the inter-model differences in variability are generally significant over the most of the analysed domain. However, some pairs of models share some similarities in the simulated total climate variability, such as the GISS-ER and GISS-EH models over land, while larger differences are found over the Hudson Bay. These differences have been attributed to the fact that the two models differ only by their ocean component.

We proposed a theoretical framework for choosing the most appropriate method for reconstructing the missing members in a multi-model ensemble. We attributed a rather positive answer to Q1 and then the SMP method can be applied to reconstruct artificial time series. The second question (Q2) results in a rather negative answer and hence the MMP method seems less appropriate for an application to this MME. It is worth noting that a positive answer to both questions would suggest the application of both the SMP and MMP methods, but in that case the MMP would provide a larger pool of climatic time periods to resample from and hence should be preferred.

We acknowledge that more complex testing frameworks could have been implemented for obtaining a more precise answer to questions Q1 and Q2. For example, the changes in the internal variability of the climate system for the next century were neglected in the present study since they are expected to be rather small for temperature over midlatitudes (Hawkins and Sutton, 2011, 2009; Räisänen, 2002) and hardly detectable due to the poor sampling of realisations for each model.

Appendix 3.A : Applying the ergodic assumption to climate model simulations

The ergodic theory has been developed through research in statistical physics (e.g. Reif 1965). As a general definition, the ergodicity principle applies when a characteristic of a system is invariant according to different coordinates (axes). In the present, we apply the ergodicity principle between the time and the "members" axes as described in the following example.

Let us first consider an AOGCM used to simulate the planetary weather over a long climatic time scale (e.g. several centuries). We assume a sufficiently long spin-up period has been removed at the beginning of the simulation (Stouffer, 2004) after which the simulation has reached some equilibrium between the main components of the model (e.g. atmosphere, ocean, land, sea-ice, vegetation). The model is run under stationary conditions, i.e. that no external transient forcings are applied (e.g. GHGA emission and land-use scenarios, volcanic emissions, variations in the solar radiation). We note that the diurnal and annual cycles in solar radiation are included, resulting in simulations that are cyclo-stationary. Let us now suppose that we generate a large ensemble of such simulations by using the same model but with slight differences in the initial conditions. After the spin-up period, the two simulations are expected to be totally uncorrelated at every time scales from day-to-day variability to longer cycles as the multi-decadal climate variability.

Suppose now that we proceed to averaging over seasonal, annual or longer time periods in order to focus on a climatic time scale, thus removing both the daily and annual cycles in the time series. The simulations being now stationary and assuming sufficiently long simulation period with a large ensemble, it is expected that the time average over one time series will tend to equal the ensemble average at any specific time. Similarly, the temporal variance calculated from one simulation (natural variability) will tend toward the inter-member variance at any time (internal variability), and similarly for the higher statistical moments.

Such an ensemble of simulations can be considered as ergodic in a similar way as used in statistical physics. According to Reif (1965), the ergodic assumption can be stated as "each system of [an] ensemble will in the course of a sufficiently long time pass through all the values accessible to it". In our example, a system consists in a realisation and the ensemble is formed by all the realisations available for a given model. If considering a sufficiently long time scale, it is expected that each realisation will visit all its accessible states but at different moment in time since the realisations are independent from one another. Corollary, the statistics calculated in time for one realisation are expected to be equal to the same statistics but calculated over all the realisations at one specific time.

It is interesting to note that for such an ensemble under ergodic conditions, the only difference between the time and member axes is the chronology of the events that characterises the time axis. On the other hand, the member axis can be seen as time axis but without any preferred order of chronology.

Appendix 3.B : Approaching stationary conditions by detrending the ensemble mean

The simulations provided by the CMIP3 multi-model (MMD) dataset include important external transient forcings and then the ergodic assumption (see Appendix 3.A) is not expected to hold. However, stationary conditions can be approached for a particular model by "correcting" its ensemble mean (i.e. the average over the realisations). If the higher statistical moments are not processed as the ensemble mean, the resulting ensemble can be seen as under "weakly stationary" conditions. One should note that a weakly stationary process generally involves both of the two first statistical moments (von Storch and Zwiers, 1999).

The ensemble mean represents the mean response of a model to its external transient forcings in the limit of a sufficiently large ensemble size. In the simulations from the CMIP3 experiment, the most important external forcing being the emission scenario, its general effect on the simulations can be described using a 4th-degree polynomial

regression as done by Hawkins and Sutton (2009). Another effect that can be described by a 4th-degree function is the secular changes in solar radiation along the 20th century. The effect of volcanic emissions and the periodic variations in the solar radiation will survive to the detrending because those happen synchronously across all the realisations. While the effect of the former is expected to be important compared to the latter for the surface air temperature, the two effects are difficult to separate using the CMIP3 archive since the models generally include both or neither of these forcings.

As an example, Fig. 3.8 shows the coefficient of determination (R^2) that characterises the fit of a 4th degree polynomial function to the ensemble mean of the GISS-ER ensemble of simulations. The coefficient of determination informs us about the proportion of variation that is described by the regression compared to the total variability about the overall mean (averaging over the time and the realisations). As seen on this figure, the values of R^2 are higher than 90% over Pacific Ocean (PO) while relatively small ($< 50\%$) over the Labrador Sea (LS). In Fig. 3.9a and b are shown the summer-mean time series over two grid points that correspond to the previous regions. In these panels, the different realisations are shown as thin coloured lines and the ensemble mean as a black line. The red line represents the polynomial function fitted to the ensemble mean. These curves show that, even for regions with small R^2, a 4th-degree polynomial function seems to describe properly the general trend detected in the ensemble mean.

As noted previously, the effects from some transient forcing agents are expected to remain in the climate simulations after the detrending of the ensemble mean. In Fig. 3.10 are shown the domain averaged (over North America) time series for the available members of each models, the ensemble means and the related polynomial fits. The GISS-ER model (Fig. 3.10d) shows three large peaks in the ensemble mean that corresponds to important volcanic events throughout the 20th century : Novarupta in 1912, El Chichón in 1982 and Mt. Pinatubo in 1991. The other models presented in the figure generally show a weaker response to volcanic forcing compared to the GISS-ER and GISS-EH models. Also, it can be seen from a general point of view that the structure of the

ensemble mean is generally different between the 20th and 21st centuries (e.g. PCM in Fig. 3.10f). In the 21st century, the ensemble mean is relatively close to the fit with relatively short cycles while longer cycles can also be seen in the 20th century. These oscillations are partly attributable to the numerous volcanic emissions that have been recorded (Ammann et al., 2003) for the 20th century and that are used to force most of the models of the MME.

Appendix 3.C Testing the ergodic assumption for a single-model ensemble

In this appendix, we describe a testing framework in order to investigate the ergodic assumption for a single-model ensemble of simulations. This test aims to provide an answer to the first scientific question ($Q1$) asked through the theoretical framework presented in Fig. 3.1.

A single-model ensemble of simulations can be represented by a matrix $\mathbf{X}$ as shown in Fig. 3.2. Using the linear model (3.1), a one-way analysis of variance (ANOVA; von Storch and Zwiers 1999) can be applied to decompose the total variability in $\mathbf{X}$ into a sum of squares, i.e. $SST = SSA + SSE$, where SST is the total sum of squares, SSA the sum of squares due to the treatment in time and SSE the residual error. These three components are summarised in Table 3.2 where the "o" notation indicates averaging over the missing subscript. Also is shown in this table the number of degrees of freedom (df) associated to each sum of square.

In order to build a test statistics that translates the ergodic assumption for the ensemble shown in Fig. 3.2, we define the null hypothesis as :

$$H_0^{ergo} : \sum_{t}^{N_T} a_t^2 = 0. \tag{3.3}$$

The null hypothesis (3.3) means that there is no treatment along the time axis and hence that all the variability in the ensemble matrix (Fig. 3.2) is described by the residual error (e_{tk}). The errors are assumed to be independent and identically distributed (iid) and

hence represent the ergodic (unforced) component of the matrix, i.e. an invariance of the statistics according to time and members. On the other hand, a_t represents the forced variability in time and hence the non-ergodic component. It is worth noting that in the context of a single-model ensemble including the GHGA forcing, H_0^{ergo} is expected to be strongly rejected.

Before constructing a test statistics for the ergodic assumption, let us consider the term SSA as shown in Tab. 3.2. Using (3.1), it can be shown that the expectation of SSA can be written as :

$$\frac{E(SSA)}{N_T - 1} = \frac{N_K \sum_t a_t^2}{N_T - 1} + \sigma_{IV}^2 \tag{3.4}$$

where

$$\sigma_{IV}^2 = \frac{E(SSE)}{N_T \times (N_K - 1)} \tag{3.5}$$

is the variance of the *iid* process (e_{tk}) which can be associated to the model simulated internal variability. From (3.4), it can be seen that $E(SSA)/(N_T - 1)$ estimates σ_{IV}^2 when H_0^{ergo} is true and a larger number if H_0^{ergo} is false. Similarly, it can be shown that $E(SSE)/(N_T \times (N_K - 1))$ estimates σ_{IV}^2 independently of whether if H_0^{ergo} is true or false. In order to test H_0^{ergo}, we then use the following ratio :

$$F_1 = \frac{SSA/(N_T - 1)}{SSE/(N_T \times (N_K - 1))}. \tag{3.6}$$

Under the null hypothesis, F_1 follows an F-distribution that is defined by its number of degrees of freedom, i.e. $F(N_T - 1; N_T \times (N_K - 1))$. The ergodic assumption can then be tested using a one-sided test at some significance level by using the critical values associated to the F-distribution. The F_1 statistics provides information about the rejection of the null hypothesis but tells very little about the physical relevance of of the non-ergodic component. The proportion of variance (P_1) of the matrix $\mathbf{X}$ that is described by the forced component of variability can be estimated as :

$$P_1 = \frac{\hat{\sigma}_F^2}{\hat{\sigma}_F^2 + \hat{\sigma}_{IV}^2} \tag{3.7}$$

where $\hat{\sigma}_F^2$ estimates the forced variance that occurs in time, which can be defined as

$$\sigma_F^2 = \frac{\sum_t a_t^2}{N_T - 1}, \tag{3.8}$$

and thus estimated by replacing (3.8) into (3.4) as :

$$\hat{\sigma}_F^2 = \frac{SSA/(N_T - 1) - SSE/(N_T \times (N_K - 1))}{N_K}. \tag{3.9}$$

The ratio (3.7) is finally calculated as :

$$P_1 = \frac{SSA - \frac{N_T - 1}{N_T \times (N_K - 1)} SSE}{SST - SSE/N_T}. \tag{3.10}$$

It is worth noting that P_1 can be written as a function of F_1 :

$$P_1 = \frac{F_1 - 1}{F_1 + (N_K - 1)}. \tag{3.11}$$

Recalling that a rejection of the null hypothesis at some significance level involves an F_1 ratio that exceeds some critical value calculated from the F-distribution, one can calculate the corresponding critical value of the variance ratio (P_1) by using (3.11). The variance ratio allows to appreciate the physical significance of the non-ergodic component when H_0^{ergo} is rejected.

Appendix 3.D Testing the inter-model differences in the simulated total climate variability

In what follows, we develop a testing framework for question *Q2* that investigates the inter-model differences in simulated total climate variability. From Tab. 3.2, the total sum of square divided by its number of degrees of freedom allows to estimate the total

climate variability (σ^2_{tot}) :

$$\hat{\sigma}^2_{tot}(m) = \frac{1}{N_K(m) \times N_T - 1} \sum_{t}^{N_T} \sum_{k}^{N_K(m)} (X_{mtk} - \overline{\overline{X_{moo}}^{t}}^{k})^2. \tag{3.12}$$

where m is the model index. Since the simulations have been detrended (Appendix 3.B), the ensemble mean X_{moo} tends to zero. Also, it is worth noting that the estimate of the total climate variability given by (3.12) tends to be equal to the sum $\hat{\sigma}^2_F + \hat{\sigma}^2_{IV}$ (see Eqs. 3.5 and 3.9) when N_T and N_K become large. We thus define the null hypothesis of equal total climate variability between two models (m and m') :

$$H_0^{var} : \sigma^2_{tot}(m) = \sigma^2_{tot}(m'). \tag{3.13}$$

This assertion can be verified through the use of an F-test defined as the following ratio :

$$F_2(m, m') = \frac{\hat{\sigma}^2_{tot}(m)}{\hat{\sigma}^2_{tot}(m')}. \tag{3.14}$$

Under the null hypothesis (i.e. when H_0^{var} is true), the F_2 ratio is distributed as the F-distribution $F\big(N_K(m) \times N_T - 1; N_K(m') \times N_T - 1\big)$. It follows that H_0^{var} can be tested by using a two-sided test at some significance level.

As a measure of physical significance, we define the relative error of variance (P_2) as :

$$P_2(m, m') = \frac{\hat{\sigma}^2_{tot}(m) - \hat{\sigma}^2_{tot}(m')}{\frac{\hat{\sigma}^2_{tot}(m) + \hat{\sigma}^2_{tot}(m')}{2}} \tag{3.15}$$

This ratio can be expressed as a function of F_2 as follows :

$$P_2 = \frac{F_2 - 1}{\frac{F_2 + 1}{2}}. \tag{3.16}$$

It results that a critical value obtained for F_2 can be converted into a corresponding critical value for P_2 and vice versa.

Tab. 3.1: Names of the models in the CMIP3 multi-model dataset that provide two or more realisations following the A1B emission scenario. Is also given the number of realisations (N_K) that are available for each model. For supplementary information, the reader is invited to refer to the PCMDI website at `http://www-pcmdi.llnl.gov`.

	Model name	$\mathbf{N_K(m)}$	**Volcanic**	**Solar**
a	CCSM3	7	x	x
b	CGCM3.1(T47)	5	-	-
c	MRI-CGCM2.3.2	5	x	x
d	GISS-ER	4	x	x
e	MPI-ECHAM5	4	-	-
f	PCM	3	x	x
g	GISS-EH	3	x	x
h	FGOALS-g1.0	3	-	-
i	MIROC3.2(medres)	3	x	x
j	ECHO-G	3	x	x
k	GISS-AOM	2	-	-

Tab. 3.2: One-way analysis of variance table where the total sum of squares, the treatment sum of squares and the sum of squared errors are expressed with their respective number of degrees of freedom. N_T is the number of time periods and N_K is the number of realisations generated using the climate model.

Component	*Sum of squares*	*Degree of freedom (df)*
total	$SST = \sum_t^{N_T} \sum_k^{N_K} (X_{tk} - \bar{X}_{oo})^2$	$N_T \times N_K - 1$
treatment	$SSA = N_K \sum_t^{N_T} (\bar{X}_{to} - \bar{X}_{oo})^2$	$N_T - 1$
error	$SSE = \sum_t^{N_T} \sum_k^{N_K} (X_{tk} - \bar{X}_{to})^2$	$N_T \times (N_K - 1)$

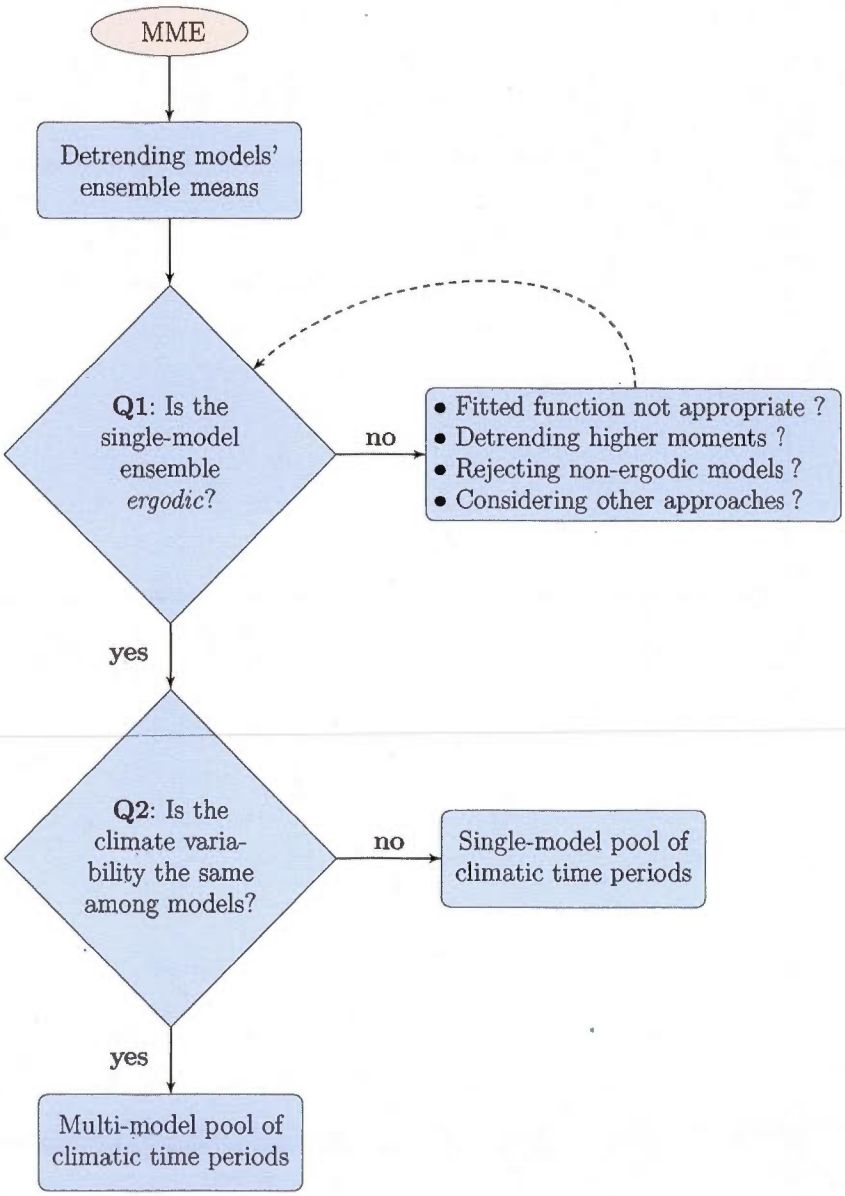


Fig. 3.1: Theoretical framework for an educated guess in the selection of a member-reconstruction method to be applied to a multi-model ensemble (MME) under transient forcing.

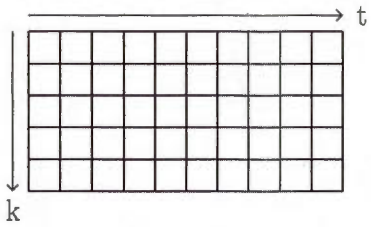


Fig. 3.2: Single-model ensemble schematised as a matrix (X) of time periods. The index t represents the N_T time periods and k represents the N_K realisations (or members) that differ in the initial conditions.

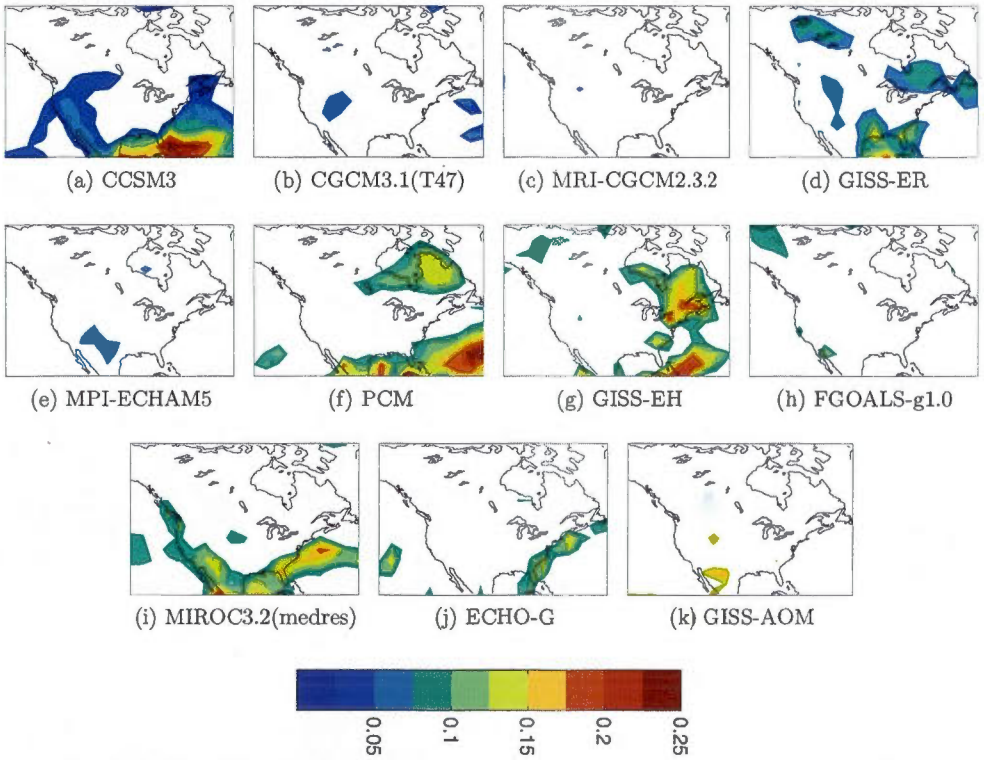


Fig. 3.3: Testing the ergodic assumption (H_0^{ergo}) using a one-sided F-test at the 10% significance level. The colored areas indicate where H_0^{ergo} is rejected over the domain. The ratio of variance (P_1, see Appendix 3.C) is shown in order to appreciate the physical significance when the ergodic assumption is rejected. The results are shown for the simulations over the 20th century with a climatic time period of 1 year ($N_T = 100$) and the models are labeled from the largest (panel a) to the smallest (panel k) single-model ensemble size (N_K) according to Tab. 3.1.

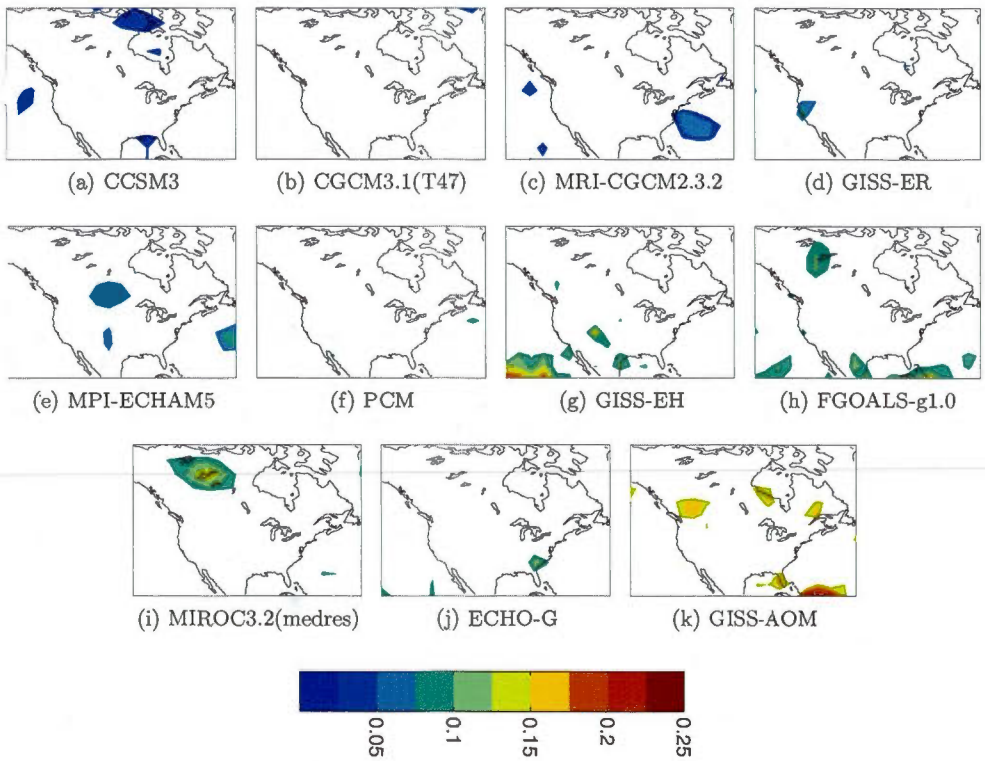


Fig. 3.4: Idem to Fig. 3.3 but for the 21st century.

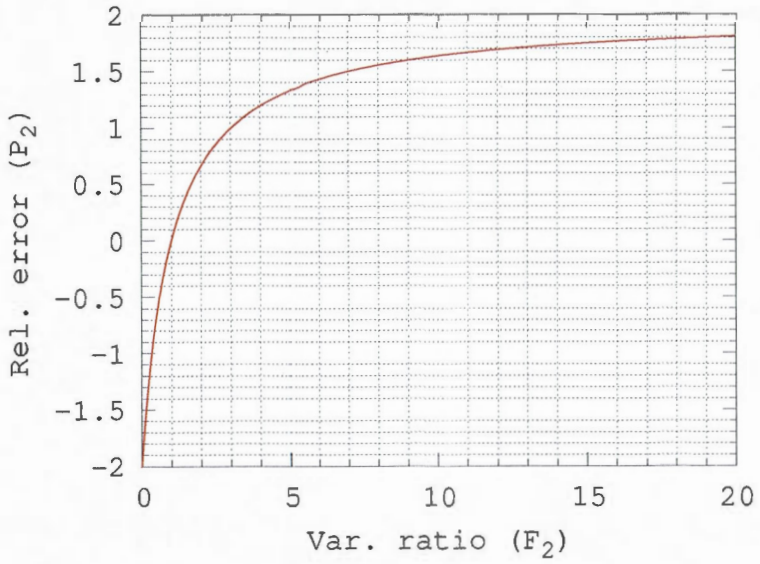


Fig. 3.5: Relative error of variance (P_2) as function of the variance ratio (F_2) of the total climate variability as simulated by two models (see Appendix 3.D).

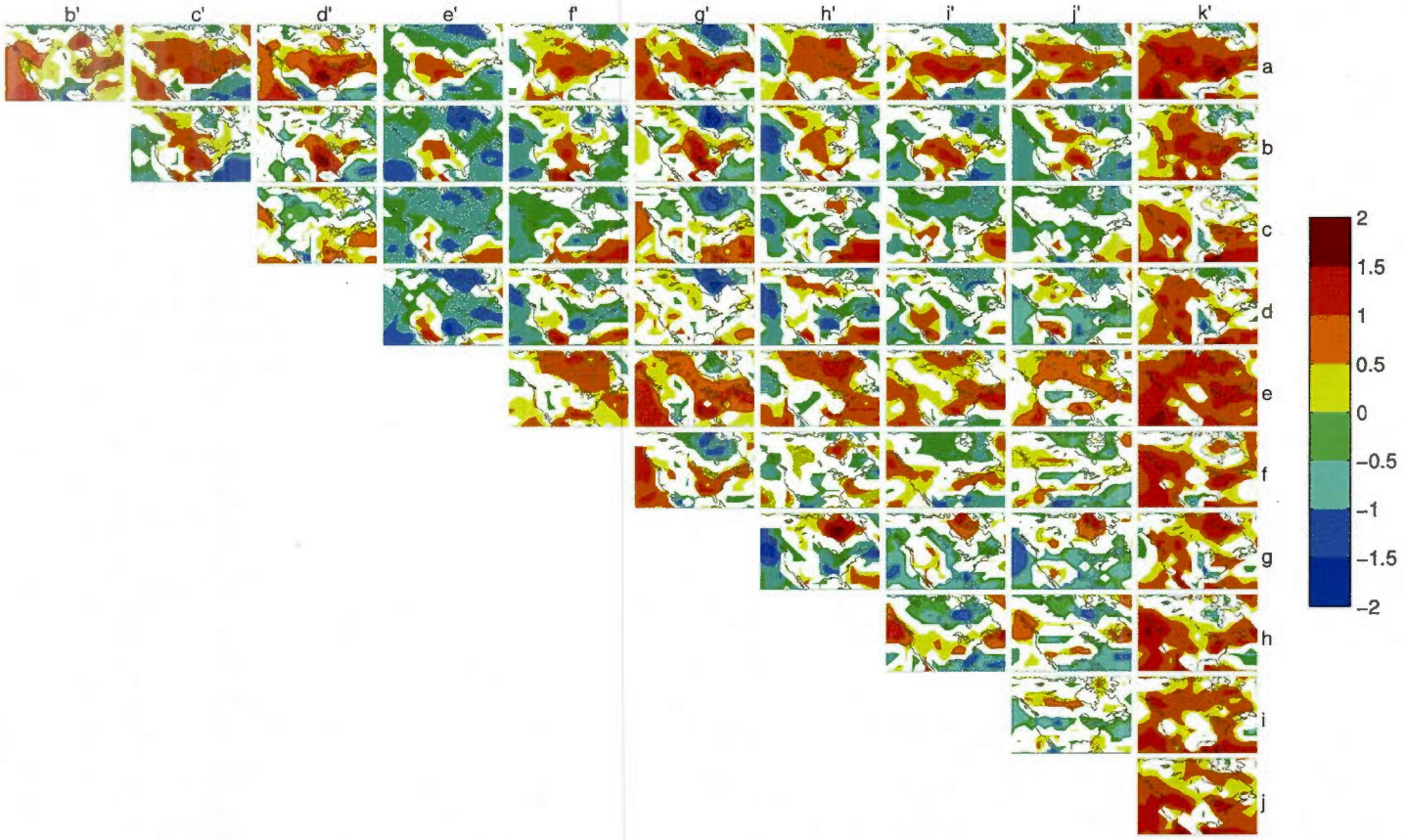


Fig. 3.6: Cross-model comparison in the simulated total climate variability over the 20th century. The comparison is based on a two-tailed F-test at the 10% significance level.

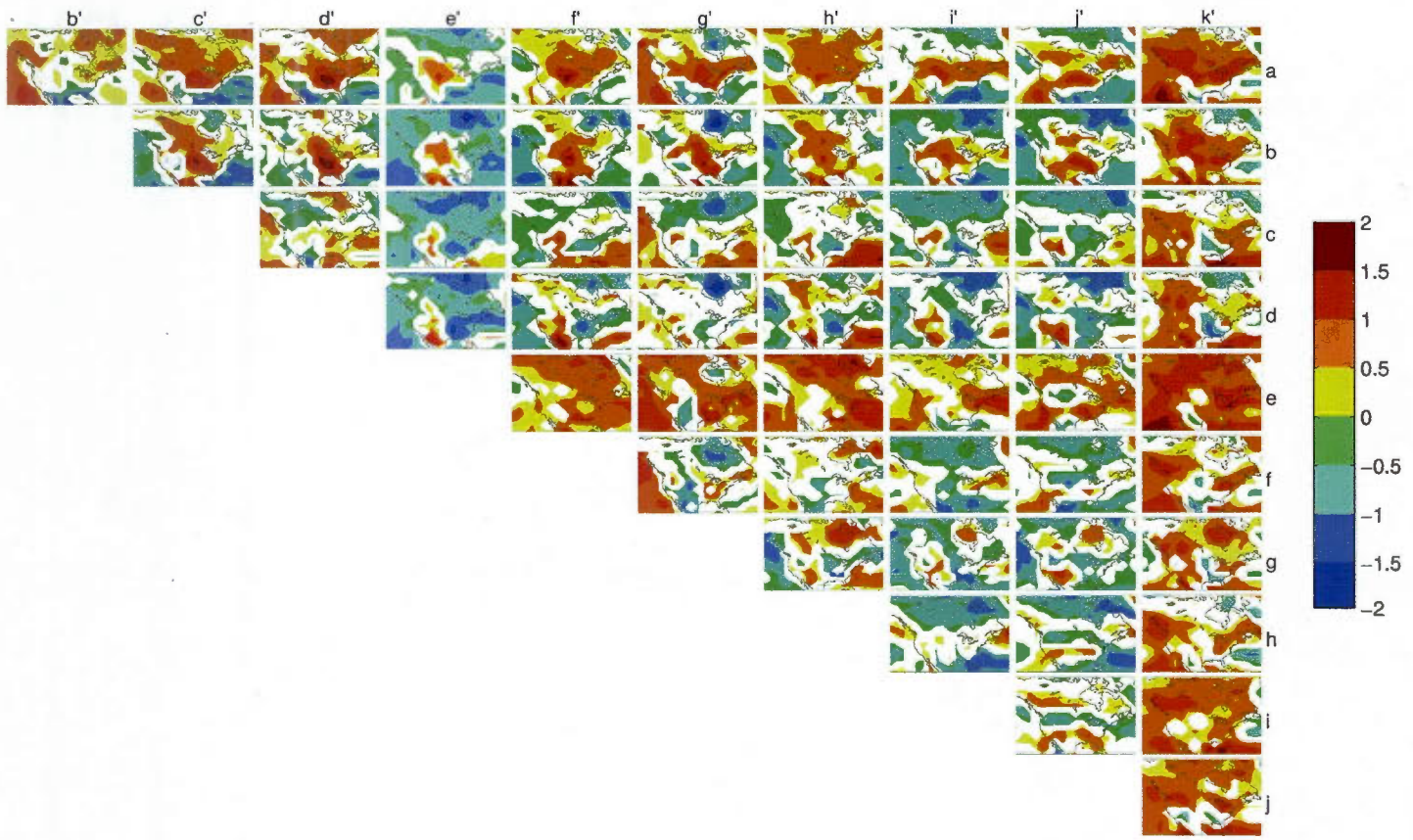


Fig. 3.7: Idem to Fig. 3.6 but for the 21st century.

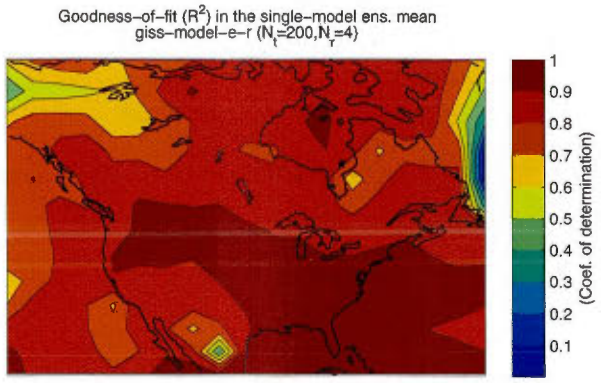


Fig. 3.8: Coefficient of determination (R^2) obtained for the fit of a 4^{th} degree polynomial function to the ensemble mean of the GISS-ER model.

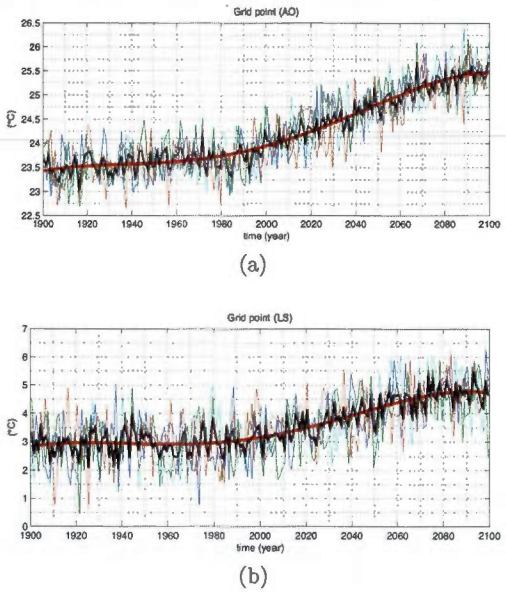
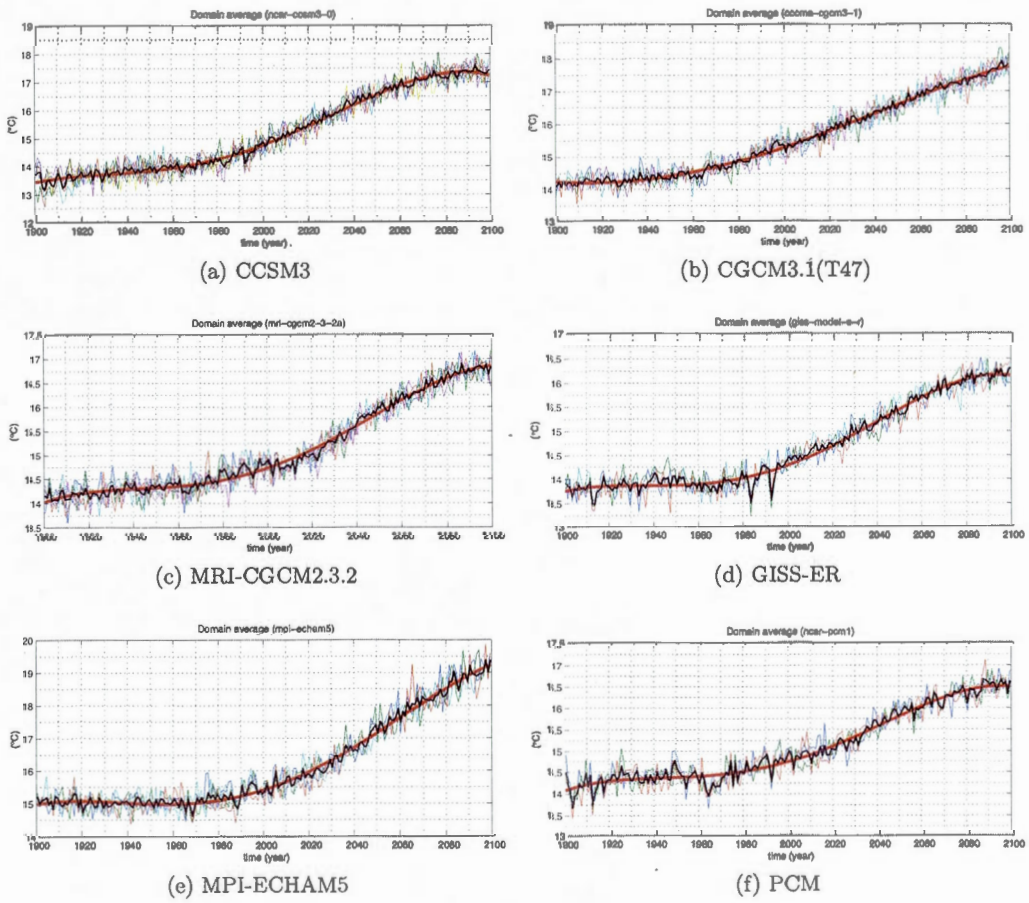


Fig. 3.9: Examples of time series for the four realisations (thin colored lines) available for the GISS-ER model. The series are shown for two grid point located over a) the Atlantic Ocean and b) the Labrador Sea. The black lines represent the ensemble mean and the red line the 4^{th} degree polynomial fit to the ensemble mean.



(a) CCSM3

(b) CGCM3.1(T47)

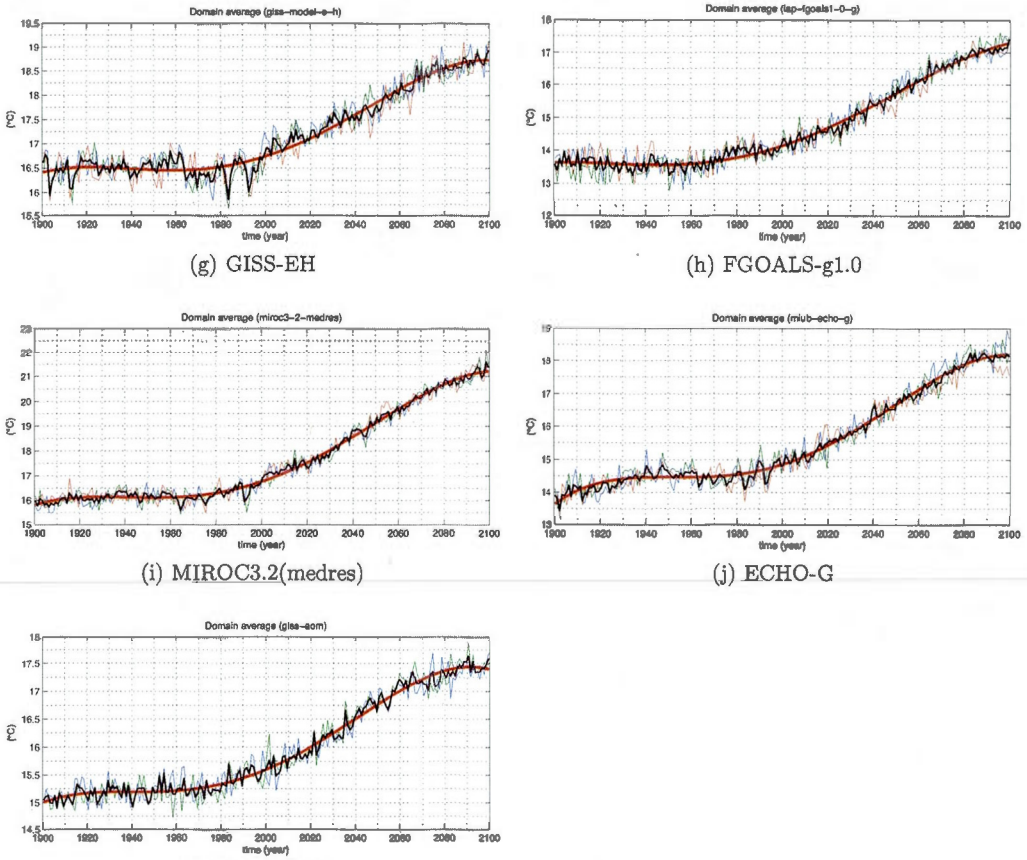
(c) MRI-CGCM2.3.2

(d) GISS-ER

(e) MPI-ECHAM5

(f) PCM

Fig. 3.10: (To be continued...)



(g) GISS-EH

(h) FGOALS-g1.0

(i) MIROC3.2(medres)

(j) ECHO-G

(k) GISS-AOM

Fig. 3.10: Domain averaged (over North-America) time series of surface air temperature covering the 1900-2100 period under the A1B scenario for the 11 AOGCMs of the multi-model ensemble (Tab. 3.1). In each panel are shown the available realisations (colored thin lines), the single-model ensemble mean (black) and the polynomial fit (thick red).

CHAPTER IV

SUMMARY AND EXAMPLES OF APPLICATION

ABSTRACT

In this last chapter, we proceed to a review of the main theoretical concepts that have been developed throughout this thesis, followed by two examples of applications. The first example compares different possible approaches for obtaining an estimate of the natural variability representative of the entire multi-model ensemble. In the second example, the "same-institute assumption" is once again investigated through an improved test statistics that focuses primarily on temporal variability of the time series rather than on the inter-member spread.

4.1 Introduction

State-of-the-art climate-change projections using Atmosphere-Ocean General Circulation Models (AOGCMs) are subject to uncertainties, which are often divided into three main components. An important one is related to the external forcings that are applied to the models, which generally consist in the emissions of Greenhouse Gases and Aerosols (GHGA) that are uncertain since based on scenarios representing the future socio-economical, technological and political context. By assuming a given pathway of GHGA emissions, another component of the uncertainty affecting the projections is known as model uncertainty, which is clearly seen by the fact that different models offer different responses to the same emissions pathways. The model uncertainty is also sometime addressed from the point of view of a single model that may exist in an arbitrary number of versions that differ in the tunings of weakly constrained parameters

(Stainforth et al., 2005; Rowlands et al., 2012). The third component is the natural variability that affects the projections of every model and that can be sampled by generating a large number of realisations with perturbed initial conditions. One could also think of two supplementary levels of uncertainty that lie below the natural variability of an AOGCM. These appear when proceeding to the downscaling of an AOGCM simulation through dynamical downscaling using Regional Climate Models (RCMs) or through statistical downscaling models (SDM) in order to obtain fine-scale details from AOGCM simulations. In such an example, a fourth level of uncertainty could be that due to the different RCMs or SDMs used for downscaling a given AOGCM realisation. A fifth level of uncertainty would be the inter-member variability of the RCM that can be sampled in the same way as for an AOGCM, that is by generating several realisations from different initial conditions.

In the last decades, several internationally coordinated projects have been conducted in order to sample the different sources of uncertainty. However, the latter sources of uncertainty are generally investigated quite differently across the projects. For example, the CMIP3 multi-model dataset (Meehl et al., 2007b) sampled several GHGA scenarios, AOGCM models and realisations thereof. At a different level, the North American Regional Climate Change Assessment Program (NARCCAP ; Mearns et al. 2009) [1] used a few AOGCMs to drive a set of RCMs under a single GHGA emission scenario. Ensembles of opportunity hence exist in a broad variety that may be seen as different attempts at assessing the main sources of uncertainty in climate-change projections.

While these ensembles provide an appreciable number and diversity of climate-change projections, these numerous pieces of information are sometime difficult to combine and to interpret. In particular, these ensembles raise important conceptual issues depending on the sampling of the different sources of uncertainty. A first concern is the unclear sampling of the models. The assumption that different climate models provide independent pieces of information about climate change is likely to be false for seve-

1. http://www.narccap.ucar.edu

ral reasons, while quantifying this lack of independence represents a fairly difficult task (Tebaldi and Knutti, 2007). One reason to believe that the climate models are not independent is that science develops based on the sharing of knowledge, for example the modelling groups learn from each other and even share parts of model code. Models are generally based on similar basic physical assumptions and include similar processes and interactions. Also, through their evaluation process, models are often tuned against the same sets of observations, what is likely to induce common biases to the models, especially since observations also contain errors. Naturally, structural similarities between climate models are likely to be strengthen when the latter are developed by nearby actors, e.g. within a same research institute that may contribute several models or model versions in large multi-model ensembles such as CMIP3.

Another characteristic common to most ensembles is that the entire matrix of all potential combinations of models and forcing is not realised since climate simulations are expensive to produce. For example, some experimental frameworks (e.g. NARCCAP) are constrained in order to minimise potential biases and statistical errors related to the incomplete sampling process. On the other hand, in unconstrained experimental frameworks such as CMIP3, the missing simulations are likely to be distributed unevenly across the ensemble; the scenarios and the realisations are sampled unevenly among the models. Such biases in the sampling process of an ensemble are also intimately related to the unequal resources and the different interests of the participating groups.

In this thesis, we noted that in a multi-model ensemble such as CMIP3, some models are represented by several realisations of a given scenario while others provide a single one. Providing at least a few realisations is important in order to obtain a climate-change signal that is more representative of a given model; averaging over multiple realisations filters noise, i.e. the natural climate variability, which might otherwise obscure some features in the signal. Several realisations of a given experiment allow assessing the natural climate variability at any point in time when the simulations are run under transient forcings.

When some elements of the matrix are missing, this may lead to situations where the experimental framework is unbalanced. For example, applying an analysis of variance (ANOVA) for decomposing the uncertainty into components (e.g. scenario, model and natural variability) on an unbalanced ensemble should involve experts' judgement in order to prevent potential biases. When a non-systematic ensemble design is not appropriate for an analysis such as ANOVA that assumes some balance in the data, qualitative methods of comparisons can be more suitable (Rowell, 2006). On the other hand, artificially correcting the imbalances of ensembles may allow performing an analysis as in an ideal case. This has been done by Déqué et al. (2007, 2012) who used data-reconstruction methods in order to transform the non-systematic framework into a systematic one and hence to apply the ANOVA. Such an approach allows simplifying the uncertainty decomposition while aiming at limiting biases and sampling errors.

Throughout this thesis, we focused on two specific sources of uncertainty in climate modelling, namely the natural variability and the inter-model spread. The first source being intrinsic to the models, the way it is quantified is very important. Another motivation for a clear quantification of this source of uncertainty is that it is of primary importance when investigating other sources, such as the model uncertainty. The two examples presented below summarise this idea. In Example 1, we take an overview of the different approaches that can be considered for combining the natural variability from an ensemble of several models. In Example 2, we choose one of the latter methods for assessing the natural variability in order to test the differences between climate-change signals simulated by different models. The choice of the method has been made in order to maximise possibilities of comparisons, including models that do not provide more than a single realisation of a specific experiment.

4.2 Theoretical summary : Review of concepts

4.2.1 Pre-selection of the simulations

In Chap. 1, we developed a framework based on resampling methods (bootstrap) in order to quantify the uncertainty of the ensemble statistics that emerges from the numerous choices available to the user when selecting a limited set of simulations from a large ensemble. While this approach allows for sampling the statistical uncertainties that emerge from a weakly constrained sampling process, we highlighted the distinction between the known sources of uncertainty (e.g. natural variability) and the extent to which these are "perceived" in the ensemble statistics.

One question that we addressed in this thesis is how the natural variability affects the ensemble statistics through the selection of a set of simulations from the large ensemble. It has been shown in Chap. 1 that the real effect of the natural variability on the ensemble statistics is underestimated due to the relatively small sample sizes of single-model ensembles (typically from 1 to 7 realisations). Such a question could be addressed in a more general way : How would the natural variability really affect the ensemble statistics given an infinite number of realisations available for each of the models ?

While the pre-selection of realisations is often done randomly (e.g. Bombardi and Carvalho 2011, Peings and Douville 2010, Räisänen et al. 2010), the selection of a set of models should preferably be constrained by some criteria. A broad range of constraints are commonly used, for example based on the same-institute criterion (Whetton et al., 2007). Ensemble post-filtering is also sometime generalised be attributing weights to the models, based on performance criteria (Giorgi and Mearns, 2002) or other physical constraints (Allen and Ingram, 2002). In the model-sampling method proposed in Chap. 1, we generalised the selection process by allowing model replacement based on the hypothesis that the CMIP3 multi-model ensemble is only a representative sample of a notional larger population of models with similar level of complexity.

4.2.2 Initial sampling of an ensemble of opportunity

As discussed in Chap. 2, suspected lacks of independence between modelling approaches is likely to be an issue when using multi-model ensembles for climate-change projections. While it is not clear to which extent it affects the climate-change projections, it involves a risk of biasing the estimated signal toward groups of similar models. Without any robust measure of model independence, attributing a higher confidence to a specific climate outcome that makes consensus between the models becomes highly questionable. Another important point is the attribution of climate-change projections uncertainty to the inter-model spread, which is rather unclear without a robust definition of model independence.

The "same-institute assumption" can be used as a cautious approach to prevent non-informative consensuses from contaminating the results of an ensemble, at the cost of reducing its size and thus increasing errors in the statistics. This can be seen as a rather conservative way of approaching ensembles and their results, while larger uncertainties are not always interpreted in a positive way by the public. On the other hand, basing adaptation and mitigation plans on overconfident results is surely not a suitable option either. The same-institute assumption consists in attributing consensuses in the models' output to some dependencies between climate models such as structural similarities. While using this criterion as a rule of thumb should involve care, it is at least very useful for pointing out groups of models that are structurally similar within a large multi-model ensemble. On the other hand, the observed propensity of models to give similar results when developed by nearby actors is probably only the tip of the iceberg concerning the more general issue of a lack of independence between the existing modelling approaches.

4.2.3 The *ergodic* assumption as a workaround for unbalanced ensemble frameworks

Large intercomparison projects are often formed in a rather open manner that favours the number and diversity of simulations over balance between experimental units. The dimension and shape of the resulting matrix of simulations is then affected by external factors such as the unequal resources and the different interests of the participating modelling centres. Hence, such ensembles may not consist in balanced designs for specific investigations such as decomposing the uncertainty into its main components. By analysing such an ensemble from the point of view of a perfect balanced framework, some elements appear to be missing, which leads to approximations in the statistical theory and possibility of biases in the results.

One way for circumventing such issues is by considering reconstruction methods in order to obtain a balanced framework from the unbalanced one and hence to facilitate the analysis by applying exact theory for analysis. Obviously, such an approach involves the risk of adding supplementary noise to the dataset. Using robust physical assumptions in the reconstruction methods is hence of primary importance in such a context. A credible physical assumption has been identified in Chap. 3 that single-model ensembles are ergodic in the sense that the temporal variability is statistically indistinguishable from that occurring between members. This characteristic of single-model ensembles is expected to occur when simulations are run under stationary conditions, while under sufficiently strong transient forcings, the ergodic assumption has to be rejected. In the latter case, however, ergodicity can be approximately reached under "artificial stationarity" achieved by detrending the time series.

As will be seen in Example 1, the ergodic assumption could be of use in the development of reconstruction methods, especially in cases where imbalance makes the analysis problematic. In Example 2, the benefits of using the ergodic assumption will be demonstrated through the construction of a test statistics that can be applied to compare two

models providing single realisations. In this case, the ergodic assumption allows reducing an important imbalance in the ensemble that consists in some models providing only a single member per experiment.

4.2.4 Analysis of variance and decomposition of the uncertainty

The analysis of variance (ANOVA) is a popular approach for decomposing variances into a number of sub-components. This approach is known to be suitable with balanced framework, or when imbalances are relatively unimportant. In order to understand how the ANOVA is affected by the imbalance in the sampling of the realisations in the CMIP3 multi-model ensemble, Appendix 4.A shows how the ANOVA can be applied to such an ensemble.

The approach is based on a statistical model of the form $X_{mn} = \mu + a_m + e_{mn}$, where X_{mn} consists in the n^{th} member available for the m^{th} model, μ the theoretical mean of the population[2], a_m the treatment effect due to the use of different models and e_{mn} the residual variability that represents the natural variability as simulated by the climate models. While in general, the latter component represents a level of noise that is independent and identically distributed (*iid*) along both m and n axes, this assumption is not expected to hold for a multi-model ensemble. Particularly, it has been shown in Chap. 3 that the natural variability is sometimes simulated rather differently across the models of the CMIP3 multi-model dataset.

Another important point is the relative importance of the different models in the calculation. In the Appendix 4.A, it is also shown how the resulting estimate of the multi-model natural variability is biased toward the models with the largest sample sizes. Another issue that is strictly due to the unbalanced design is that the component of inter-model variance cannot be obtained explicitly, but only approximately. This is due to the fact

---

2. Since defining such a population is problematic (see Chap. 1), we assume the ensemble to be representative of a larger population that includes other possible modelling approaches with a similar level of complexity as the CMIP3 models.

that we need to assume an effective sample size (N_0) in (4.20) that should represent the different models' sample sizes. Hence, when sample sizes differ importantly, it increases the range of possible values that can be taken by N_0, leading to a larger estimation error in the inter-model spread.

4.3 Example 1 : Multi-model combination of the simulated natural variability

In the context of assessing the different components of uncertainty in climate-change projections, the natural variability is the first level of uncertainty that should be estimated, while the second is the model uncertainty. Natural variability represents a measure of noise from which the physical and statistical significance of the latter can be assessed. The ANOVA is a conventional technique for decomposing the variability into its several components. However, such an approach is based on hypotheses that do not necessarily hold for an ensemble of opportunity such as CMIP3. An important assumption is that the "noise" is *iid* according to the model and member axes. As shown in Chap. 3, there is compelling evidence that the natural variability is not identically distributed across models, hence violating the assumption. Another problem related to the use of the ANOVA in that context is that the number of members largely differs across the models. As will be shown below, the unequal sample sizes and the non-identically distributed natural variability across the models are important factors to consider when estimating the primary source of uncertainty in multi-model ensembles. Also, it appears necessary to compare different approaches in order to optimise the estimate of the natural variability depending on the ensemble under consideration.

As a starting point, the member-sampling approach described in Chap. 1 allows to quantify how the ensemble statistics are affected by a random selection of one member per model. The uncertainty of the ensemble mean signal (U^{Δ}_{mem}) consists in a manifestation of the natural variability as simulated by different models. As shown in Appendix 4.B, this measure of uncertainty can be transformed into an estimate of the natural variability by "scaling" this error of the mean to a single model by using the standard error

relationship as $\sqrt{M} \times U^{\Delta}_{mem}$, M being the number of models in the ensemble. While this treatment has been applied to climate-change projections, i.e. differences between two climatic states, we use in the following the corresponding estimate of the natural variability.

Let us rewrite (4.28) from Appendix 4.B as

$$\hat{\sigma}^2_{mem}(t) = \frac{1}{M}\sum_{m}^{M}\frac{1}{N_m}\sum_{n}^{N_m}(X_{mnt} - \overline{X}_{mot})^2 \tag{4.1}$$

where the "o" notation indicates averaging over the missing subscript. X_{mnt} is not a delta but the climate state at time t of the n^{th} realisation over N_m members provided by the m^{th} model. (4.1) consists in the analytic form of the member-sampling uncertainty. This variance is calculated at a particular time (t) and consists in a multi-model average of biased estimates of inter-member variances. While such a bias is more important at small N_m, this estimate of the multi-model natural variability is not weighted according to the models' different sample sizes.

As a more conventional approach, the ANOVA allows to decompose the total variability of a multi-model ensemble in two components : the inter-model spread and the natural variability. As shown in Eq. 4.16 (Appendix 4.A), applying the ANOVA to a multi-model ensemble leads to an estimate of the multi-model natural variability that corresponds to

$$\hat{\sigma}^2_{WI}(t) = \frac{1}{N-M}\sum_{m}^{M}\sum_{n}^{N_m}(X_{mnt} - \overline{X}_{mot})^2, \tag{4.2}$$

where $N = \sum_m N_m$ is the total number of simulations in the ensemble. (4.2) consists in a combination of several models inter-member variability. By gathering the deviations from different models and dividing by the number of degrees of freedom (df), this estimate is weighted according to the number of members provided by each model and hence biased toward the models providing the largest sample sizes. Given potentially important differences in the natural variability simulated by the models (Chap. 3), it is relevant to

consider an unweighted version of (4.2) such as

$$\hat{\sigma}^2_{UI}(t) = \frac{1}{M}\sum_{m}^{M}\frac{1}{N_m - 1}\sum_{n}^{N_m}(X_{mnt} - \overline{X}_{mot})^2, \tag{4.3}$$

which gives an equal weight to each model in the resulting multi-model estimate of the natural variability. It can be seen from the latter equation that $\hat{\sigma}^2_{UI}(t)$ consists in an average over several unbiased estimates of inter-member variances.

The variance estimates (4.1) to (4.3) consist in different approaches for combining the inter-member variability from several models and these estimates pertain to a given time. Under the assumption that the inter-member spread does not change significantly with time, it is convenient to consider the information from the entire time series in our multi-model estimate of the natural variability.

Recalling (4.2), the squared deviations can be summed by including all the time periods from each of the models, and hence dividing by $T(N - M)$ the number of df, we obtain

$$\hat{\sigma}^2_{WTI} = \frac{1}{T(N-M)}\sum_{m}^{M}\sum_{n}^{N_m}\sum_{t}^{T}(X_{mnt} - \overline{X}_{mot})^2. \tag{4.4}$$

The latter time-averaged estimate of the multi-model natural variability is weighted according to the sample sizes. Similarly, we can define an unweighted version of (4.4) by summing (4.3) over time such as

$$\hat{\sigma}^2_{UTI} = \frac{1}{M}\sum_{m}^{M}\frac{1}{T(N_m - 1)}\sum_{n}^{N_m}\sum_{t}^{T}(X_{mnt} - \overline{X}_{mot})^2. \tag{4.5}$$

Finally, assuming ergodicity in the single-model ensembles, one can imagine two additional ways of combining the natural variability from several models, namely the weighted ergodic variance

$$\hat{\sigma}^2_{WE} = \frac{1}{N}\sum_{m}^{M}\sum_{n}^{N_m}\frac{1}{T - K - 1}\sum_{t}^{T}(X_{mnt} - \tilde{\mu}_{mnt})^2 \tag{4.6}$$

and the unweighted ergodic variance

$$\hat{\sigma}^2_{UE} = \frac{1}{M}\sum_{m}^{M}\frac{1}{N_m}\sum_{n}^{N_m}\frac{1}{T-K-1}\sum_{t}^{T}(X_{mnt}-\tilde{\mu}_{mnt})^2. \tag{4.7}$$

The main difference between (4.4)-(4.5) and (4.6)-(4.7) is that the former consist in multi-model time-averaged inter-member spreads relatively to single-model ensemble means ($\overline{X}_{mot}$), while the latter consider the variability of the time series relatively to a trend denoted by $\tilde{\mu}_{mnt}$. The term $T-K-1$ consists in the number of *df* associated with the mean squared error of a realisation around the K^{th}-degree trend, for which $K+1$ parameters have to be estimated. Recalling the ergodic assumption, the temporal variability around the trend includes the non-ergodic part of the signal. On the other hand, when considering the spread around a single-model ensemble mean such as in (4.4) and (4.5), the non-ergodic part of the signal is included in the mean and hence does not contribute to the final estimate of the natural variability.

4.3.1 Results

Let us now take a look to some results related to the previous estimates of the natural variability. A common feature appearing in most of the previous estimators (with exception of Eqs. 4.6 and 4.7) is that the variability emerges from deviations about single-model ensemble means. We hence limit the following investigation to the models from CMIP3 (A1B) that provide at least two realisations. As seen from Tab. 1.1, the pre-selected ensemble consists in 42 simulations from 11 models.

In Fig. 4.1a to c are presented three approaches for assessing the inter-member variability of the climate-change signal from a multi-model ensemble. Fig. 4.1a and b show the ANOVA coefficient ($\hat{\sigma}_{WI}$) and its unweighted version ($\hat{\sigma}_{UI}$) respectively, for three 20-year averaging windows (2000-2020, 2040-2060 and 2080-2100). Comparing these two approaches for each of the time periods, some differences appear in intensity but the general shape of the patterns remains similar. By comparing with 4.1c, i.e. the member-

sampling coefficient $\hat{\sigma}_{mem}$, it can be seen how the intensity of the inter-member variability is systematically smaller for the member-sampling approach than for the ANOVA coefficient and its unweighted version (Fig. 4.1a and b). This systematic bias of the member-sampling approach is mainly due to the relatively small number of members provided by the models, as can be understood from (4.1); $\hat{\sigma}_{mem}$ being a multi-model average of several biased estimates of variance.

It is possible to unbiase the member-sampling estimate of the multi-model inter-member spread by using a correction factor (G) such as $G \times \hat{\sigma}_{mem}$. Based on the approach detailed in Appendix 1.A, the "perfect" ensemble of simulations is defined as consisting in 11 models, each one being represented by 1000 realisations. On the other hand, the "imperfect" ensemble has the same structure as the present 11-model ensemble. By using Monte-Carlo methods, the perfect and imperfect ensembles are generated 2000 times, where for each ensemble, 2000 iterations of the member-sampling approach are applied. By determining the most probable value of the correction factor, the imperfect ensemble has to be inflated by $G = 1.19$ in order to suppress its bias compared to the perfect case.

In Fig. 4.1d is shown the empirically unbiased estimator of the multi-model inter-member variability ($G \times \hat{\sigma}_{mem}$). Compared to $\hat{\sigma}_{WI}$ and $\hat{\sigma}_{UI}$ (Fig. 4.1a and b respectively), patterns are now quite similar in both their shape and intensity. Moreover, $G \times \hat{\sigma}_{mem}$ appears more similar to $\hat{\sigma}_{UI}$ than $\hat{\sigma}_{WI}$, which is due to the fact that both $\hat{\sigma}_{mem}$ and $\hat{\sigma}_{UI}$ are unweighted estimators according to the number of realisations per model. By paying attention to the temporal evolution of the multi-model inter-member variability, patterns display a relatively important variability. These changes in time are mainly due to the poor sampling of members rather than to real physical changes in the natural variability, for instance due to the external transient forcings (e.g. GHGA). Also, it is worth noting that the differences between weighted and unweighted estimators appear smaller than the temporal variability of the inter-member spread.

In Fig. 4.2a and b is shown the estimates of inter-member variability obtained using the member-sampling and the ANOVA methods respectively, both being applied to a

reconstructed ensemble that has been filled up to 100 members by applying the single-model pooling method (SMP ; see Chap. 3) to each of the models. Briefly, this application of the SMP method consists in a random sampling over a pool of time periods obtained from detrended time series. Only two coefficients are shown here by simplicity, $\hat{\sigma}_{mem}(t)$ and $\hat{\sigma}_{WI}(t)$, referring to (4.1) and (4.2) respectively. Due to the nature of the SMP method, a single-model ensemble is forced to become strongly ergodic as its size increases. Hence, $\hat{\sigma}_{mem}(t)$ and $\hat{\sigma}_{WI}(t)$ tend to be constant in time for a sufficiently large number of reconstructed members. The unweighted coefficient ($\hat{\sigma}_{UI}$) is not shown here since the weighting does not have any influence when the models' sample sizes are equal. It can be seen from the results that for such a large ensemble, both the member-sampling and the ANOVA lead to practically identical results.

In Fig. 4.3a and b is shown the inter-member variability averaged over time using the weighted and unweighted coefficients ($\hat{\sigma}_{WTI}$ and $\hat{\sigma}_{UTI}$ respectively). These approaches lead to rather similar results while differences between the time-dependent versions of these estimators (Fig. 4.1a and b) were more important. This particularity is related to the error of estimation according to the sample size used in these calculations. Taking example with the weighted estimators, the sample used in the calculation of $\hat{\sigma}_{WI}$ corresponds to N deviations from M means ($N - M$ df) while $T(N - M)$ df are used in the calculation of $\hat{\sigma}_{WTI}$.

Fig. 4.4a and b shows the weighted and unweighted ergodic variances, respectively. Under the ergodic assumption, the temporal variability of the times series is calculated according to a trend (calculated separately for each model, realisation and grid point) and hence accounts for some additional variability that is due to the external forcings (mainly the effect of the volcanoes in the 20th century) that occur synchronously between the members of a same model. As can be seen by comparing Fig. 4.4 to Fig. 4.3, the ergodic variances are slightly larger than the time-averaged inter-member spreads. Also, the shape of the pattern of the weighted ergodic variance (Fig. 4.4a) is similar to the corresponding weighted time-averaged inter-member spread (Fig. 4.3a) and similarly for

the unweighted versions (Fig. 4.4b and Fig. 4.3b).

Let us now summarise the previous calculations. Fig. 4.5a presents histograms for three time-dependent ratios : $\hat{\sigma}_{WI}(t)/\hat{\sigma}_{UI}(t)$ (in blue), $\hat{\sigma}_{mem}(t)/\hat{\sigma}_{UI}(t)$ (in red) and $G \times \hat{\sigma}_{mem}(t)/\hat{\sigma}_{UI}(t)$ (in green). These are calculated for all grid points of the domain and for each 20-year averaging window from 1900 to 2100. The blue curve represents the effect of whether using a weighted or unweighted estimate of the inter-member spread. The distribution is centred on 1, which means that weighting the models may increase or decrease the estimate of the variability in a balanced manner. Another way to interpret this is that there is no correlation between the sample size and the inter-member spread, which are obviously two unrelated quantities. The red curve represents the estimate of the inter-member spread using the member-sampling approach, and in green, its empirically corrected version. By applying the correction factor, the green distribution is relatively well centred on 1, which validate the method detailed in Appendix 1.A for calculating G. The red and green distributions have rather small width, which relates the high similarity between $\hat{\sigma}_{mem}$ and $\hat{\sigma}_{UI}$ since both are unweighted estimators according to the models' sample sizes.

In Fig. 4.5b, the ratios for the estimators of the inter-member spread averaged in time are shown. Represented by the red curve, the ratio $\hat{\sigma}_{WTI}/\hat{\sigma}_{UTI}$ is centred on 1, as was the blue curve in Fig. 4.5a. The blue and green curves represents similar ratios for the weighted and unweighted ergodic variances, namely $\hat{\sigma}_{WE}/\hat{\sigma}_{UTI}$ and $\hat{\sigma}_{UE}/\hat{\sigma}_{UTI}$, respectively. These ratios are not centred at 1, which relates the supplementary contribution of the non-ergodic part of the variability that is accounted in these estimates. A maximum value is obtained at 1.1, which could be interpreted as if the non-ergodic component typically increases the inter-member spread of about 10%. Since this supplementary amount of variability is mainly due to a natural factor (volcanic emissions), determining whether the ergodic variance consists in a more realistic representation of the natural variability of the climate system depends on the definition of what should be included in that system. Assuming that these emissions are part of the climate system, the ergodic

variability should consist in a more realistic estimate of the natural variability, while the inter-member spread is simply the internal variability of the system given a set of external boundary conditions.

The grey curve gives the ratio $\hat{\sigma}_{WI}(t)/\sigma_{UTI}$, i.e. a time-dependent inter-member variability compared to a time-averaged estimator. It gives an idea of the temporal variability of the inter-member spread around its time-averaged value. As noted before, the temporal variability of the inter-member spread (grey curve in Fig. 4.5b) is more important than the variability due to the weighting procedure (blue curve in Fig. 4.5a, note the different scales). Finally, the black curve represents the ratio between the inter-member spread calculated from a reconstructed ensemble (based on the SMP method) compared to the unweighted time-averaged inter-member spread. The distribution shows that $\hat{\sigma}_{SMP}$ underestimates $\hat{\sigma}_{UTI}$, which is probably due to an oversampling of the members. While using 100 members in the SMP method gives robust results according to time, the maximum number of reconstructed members should be 20 and 70 for models providing 2 and 7 members respectively.

In Sect. 4.4, the ergodic assumption will be considered in a different manner. It will be shown how to systematically transfer the information about the natural variability between members and temporal axes, rather than by artificially reconstructing new realisations to an ensemble as with the SMP method. This approach will be applied in the context of assessing the statistical significance of the difference between climate-change signals as simulated by two models developed within a same research institute.

4.4 Example 2 : Improving statistical testing of the same-institute assumption based on ergodicity in single-model ensembles

In Chap. 2, the same-institute criterion has been applied in order to focus on specific pairs of models that share structural similarities. The "same-institute assumption" consists in relating these structural similarities to potential consensuses in the models' output. In order to compare a pair of models developed by a same institute, the difference

in the climate-change signals has been investigated by using a t-test. This test assesses the statistical significance of inter-model differences based on a measure of the natural variability. While sample sizes are relatively small for assessing the inter-member spread, a supplementary limitation to this investigation is that some pairs of models had to be excluded when only one realisation was available for a model. The CMIP3 multi-model ensemble providing a large diversity of simulations, it also consists in a great opportunity for studying potential structural similarities between climate models. While defining a robust metric of model independence is a rather complex task, if even possible, investigating typical structural differences between climate models should at least increase our knowledge about what should resemble such a hypothetical metric.

As discussed before, an important characteristic of AOGCM simulations that can be used to circumvent the limitations of single-model ensembles is that they can be assumed as ergodic, i.e. that the natural variability calculated over time is statistically indistinguishable from the inter-member spread. While one should expect perfect ergodicity in an ensemble of simulations run under stationary conditions, applying a strong external forcing (e.g. GHGA emissions) results in a violation of the ergodic assumption. As a workaround for simulations run under transient forcings, it has been shown that by detrending the time series with a 4^{th}-degree polynomial function, the remaining forced variability (e.g. volcanic emissions) is relatively small ; single-model ensembles can then reasonably be treated as ergodic under such conditions of artificial stationarity.

By imposing such artificial condition of ergodicity, our previous investigation of the same-institute assumption can be extended to the pairs of models that provide only single realisations. Rather than assessing the natural variability by using the inter-member spread (see Sect. 4.3), the natural variability is now calculated from the detrended time series of single realisations, which are then pooled together when several are available. In the following, we show how to test the difference between two climate-change signals, independently of how many members are provided for each model.

Let $\hat{\mu}_X$ and $\hat{\mu}_Y$ be estimates of the ensemble mean climate-change signals from two

models, denoted with the indices X and Y respectively. Recognising that the simulated natural variability is not equal among the models (Chap. 3), the t-statistics for the difference of the means consists in

$$t = \frac{\hat{\mu}_X - \hat{\mu}_Y}{\sqrt{\frac{\hat{\sigma}_X^2}{N_X} + \frac{\hat{\sigma}_Y^2}{N_Y}}} \tag{4.8}$$

where N_X and N_Y are the sample sizes (number of members) used in the calculation of $\hat{\mu}_X$ and $\hat{\mu}_Y$, and similarly, $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$ are the natural variability associated with the two models' climate-change signals respectively. Particularly, both $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$ consist in variance estimates related to a difference between two climatic states. For simplicity, let us define the climate change as the difference between two 20-year averaging windows (e.g. 2020-2040 relative to 1980-2000). Assuming independence of the details of future and past climatic states as a result of natural variability, the variance of the difference between two climatic states ($\hat{\sigma}_X^2$) is equal to the sum of the variances related to each of these states, i.e. $\hat{\sigma}_X^2 = 2\hat{\sigma}_{X20}^2$ where $\hat{\sigma}_{X20}^2$ is the natural variability of a time series formed by 20-year climate periods. These estimates of the natural variability can be assessed based on the application of the ergodic principle. In clear, the n^{th} realisation of the m^{th} model (X_{mnt}) has its trend ($\tilde{\mu}_{mnt}$) removed and the residual mean squared error consists in an estimate of the natural variability such as

$$\hat{\sigma}_m^2 = \frac{2}{N_m} \sum_n^{N_m} \frac{1}{T-K-1} \sum_t^T (X_{mnt} - \tilde{\mu}_{mnt})^2, \tag{4.9}$$

which consists in one particular model considered in the unweighted multi-model average of the ergodic variance as shown in (4.7) and multiplied by 2 since it is the variance of a difference. In (4.9), $K = 4$ is the degree of the polynomial function related to $\tilde{\mu}_{mnt}$ and $T - K - 1$ the number of degrees of freedom associated to the mean squared error according to the trend. The natural variability of each of the N_m members from the m^{th} model are hence averaged by summing over n and dividing by N_m.

Recalling (4.8), the denominator is not proportional to the χ^2 distribution due to the fact

that $\hat{\sigma}_X^2 \neq \hat{\sigma}_Y^2$, and hence this statistic is not t-distributed even under a true null hypothesis of equal means. Well known as the Behrens-Fisher problem, a convenient solution is to assume a t-distribution with its number of degrees of freedom being estimated from the data as

$$df = \frac{\left(\hat{\sigma}_X^2/N_X + \hat{\sigma}_Y^2/N_Y\right)^2}{\frac{\left(\hat{\sigma}_X^2/N_X\right)^2}{N_X(T-K-1)} + \frac{\left(\hat{\sigma}_Y^2/N_Y\right)^2}{N_Y(T-K-1)}}, \tag{4.10}$$

that has been constructed by using the Welch's approximate t-solution as described by Scheffé (1970). (4.10) is bounded as $min(N_m)(T-K-1) \leq df \leq (N_X+N_Y)(T-K-1)$. By comparison with our previous t-test based only on the several members without temporal averaging (Eq. 2.5 in Appendix 2.A), the approximate number of degrees of freedom associated with this t-distribution is bounded as $min(N_m) - 1 \leq df \leq N_X + N_Y - 2$. For example, by using $T = 10$, $K = 4$ and $N_X = N_Y = 2$, $1 \leq df \leq 2$ for the approach based on the multiple members (Eq. 2.5) and $5 \leq df \leq 20$ for the case using the ergodic assumption (Eq. 4.10). This important increase in the number of degrees of freedom for such a minimalist case of two members per model results in a test with a higher power to reject the null hypothesis.

4.4.1 Results

In Chap. 2, the same-institute assumption has been investigated using six pairs of models that allow a t-test by providing multiple members for at least one of the two models (Tab. 2.3). Hence, among the nine pairs of models shown in Tab. 2.2, three had to be excluded from the analysis, i.e. those related to the CSIRO, GFDL and UKMO modelling groups. The previously explained approach based on the ergodic assumption allows extending our investigation to the three excluded pairs of models in addition to increasing the power of the statistical test.

In Fig. 4.6a to f, the pairs of models already investigated in Chap. 2 are shown while in Fig. 4.6g to i, the excluded pairs providing a single realisation per model are shown. In the following, we investigate the climate-change differences by focusing on the signal (in

surface air temperature) relative to the reference period of 1980-2000. Fig. 4.6a and b shows the pairs of models that differ from the point of view of their resolution, namely the CGCM and MIROC models respectively. Recalling that the model output has been interpolated over a coarser-resolution grid, the CGCM models show very low rejection rates and hence these two models provide climate-change signals that are not statistically different for 2000-2020 and 2040-2060, while subtle differences begin to emerge at 2080-2100. For the MIROC pair, where the change in resolution is more important, significant differences emerge in 2040-2060 and fill the oceanic part of the domain by 2080-2100. It is worth noting that the high-resolution MIROC model has a larger climate-change signal than the low-resolution version over the ocean. It is also true for the CGCM pair but to a smaller extent (Fig. 4.6a).

The three GISS models are interesting to compare since differing in their atmospheric and oceanic components. Fig. 4.6c shows the difference between EH and ER models that have different ocean components. Significant differences in the sensitivity mainly occur over the Hudson Bay, where lies a difference of about 3°C in their climate-change signal in 2080-2100. The AOM and ER models (Fig. 4.6d) differ in all their components with different versions of their ocean component; the climate-change signal for AOM is generally smaller over land regions and larger over the Hudson Bay relatively to ER. These differences also increase with time, which indicates different climate sensitivities. For the AOM-EH pair (Fig. 4.6e) that consists in a difference in all the main model components, a negative minimum of difference becomes more intense with time over the western part of North America and the Hudson Bay, while there is a significant positive difference occurring over the Pacific Ocean for 2040-2060 and 2080-2100. Considering now the NCAR models (Fig. 4.6f), CCSM3 warms significantly faster than PCM from the 2000-2020 time period.

Let us now take a look to the pairs of models providing only single members, from the CSIRO, GFDL and UKMO modelling groups (Fig. 4.6g to i respectively). Versions 3.0 and 3.5 of the CSIRO model (Fig. 4.6g) differ in their parameterization (e.g. ocean

eddy transport coefficient). Important differences occur around the Hudson Bay where the climate-change signal differs of about 4°C in 2080-2100. On the other hand, these models do not show significant differences in the signal over an important part of the land area at the centre of the domain, even by the end of the 21st century (2080-2100). Versions 2.0 and 2.1 of the GFDL model (Fig. 4.6h) has structural differences that can be understood as minor modifications to the code (e.g. numerical scheme). They show practically no significant differences in their climate-change signal. Finally, the HadCM3 and HadGEM1 models from the Hadley Centre are compared in Fig. 4.6i. It is interesting to note that these two models, which are generally thought as *a priori* independent models developed within a same institute, show relatively low rejection rates of the null hypothesis of equal means. Even for the 2080-2100 period, the two climate-change signals are not significantly different from each other nearly over all the continental region. Over the Pacific Ocean and the West Coast of North America, a small area of negative difference slowly increases in magnitude with some statistical significance in 2080-2100.

4.5 Conclusions

A generally accepted idea among the community is that the three main sources of uncertainty affecting the climate-change projections consist in the natural variability of the climate system, the model uncertainty and the GHGA scenario uncertainty. The former source being related to the chaotic nature of the climate system, it would even affect the simulations from a perfect climate model. Each model consisting in an approximation of the true climate system, the second source of uncertainty represents the differences in the results related to the use of different approaches to climate modelling. The third source of uncertainty is sometimes considered as outside from the physical climate issue since GHGA emissions largely depend on socio-economical, technological and political issues. While the latter sources of uncertainty are often sampled, analysed and discussed using a variety of ensemble structures, fundamental issues remain in their interpretation and quantification. Throughout this thesis, we have focused on the two first sources,

that is the natural variability and model uncertainty.

Basically, quantifying the natural variability in climate-model simulations can be done in two different ways. Applied to a given model, a basic way of addressing this source of uncertainty is by using a single but very long climate simulation run under stationary conditions (i.e. without external forcing change), from which the temporal variability from the mean climate can be estimated. Another way for quantifying the natural variability of a climate model is by using ensembles of multiple realisations differing only in their initial conditions. The spread between the ensemble members may hence be used as a measure of the natural variability.

In large model inter-comparison projects such as CMIP3, the focus is generally on simulations run with transient GHGA emission scenarios. However, since running climate simulations is expensive in terms of time and computational resources, few realisations are generally provided by the modelling centres. Overall, the explicit sampling of the natural variability in contemporary ensembles is rather poor ; in addition it is heterogeneously sampled across the different models. For example, in the CMIP3 (A1B), more than half of the models are represented by a single realisation while the maximum sample size is 7 members, a rather large ensemble from the point of view of computational cost and data volume, but rather small in statistical sense.

In a multi-model ensemble, the imbalance of sample sizes between models firstly complicates the analysis of natural variability. Because natural variability differs between models, a multi-model combination of the simulated natural variability is beyond the scope of common statistical methods. For example, using an analysis of variance (ANOVA) as a way to estimate the inter-member variance involves the assumption that the natural variability is *iid* across members and models, which is known to be false. While in the case of equal variability, the imbalance in sample sizes leads to approximations in the ANOVA theory, the occurrence of both the imbalance and the unequal variances tends to bias a multi-model estimate of the natural variability. More clearly, the resulting estimate is weighted according to the sample size and hence is necessarily biased toward

models that are better represented in the ensemble.

In this work, we have summarised different alternatives for calculating the natural variability related to a multi-model ensemble. We presented three estimators based on the inter-member spread that apply to a specific time. The first followed the member-sampling approach detailed in Chap. 1, aiming at quantifying the effect of a random selection of a single member per model when several are available. This estimator has been shown to be a multi-model average of biased estimates of variance, unweighted, giving the same relative importance to each of the models. The second estimator is the mean squared error term related to the ANOVA. It is basically an unbiased estimator of variance under the assumption that the natural variabilities are equal. Since it is not actually the case, the ANOVA coefficient tends to be biased toward some of the models since weighted according to the sample sizes. The third estimator of the inter-member spread consists in an unweighted version of the ANOVA coefficient that gives equal weight to the models no matter on how many realisations are available. The ANOVA coefficient and its unweighted version have been also calculated by considering the entire time series, i.e. as multi-model time-averages of the inter-member spread. A supplementary pair of weighted/unweighted coefficients has been also provided, namely the ergodic variances that are based on the premise that single-model ensembles are ergodic. Unlike the time-averages of inter-member spreads, the latter estimates focus on the variability that appears in the time series once the trend is removed.

The results show that the inter-member spread varies with time, mainly due to the limited number of realisations rather than any real physical changes in the natural variability that could be attributed to changes in the external forcings (GHGA). For the estimate related to the member-sampling approach, the systematic bias has been successfully removed by applying a correction factor that depends on the structure of the multi-model ensemble and that has been obtained empirically from Monte-Carlo simulations. The difference between weighted and unweighted statistics appeared a little smaller than the temporal variability of the inter-member spread. Overall, the small

influence of the weighting could be explained by the important inter-model differences in the simulated natural variability that are filtered through the averaging procedure while the remaining weak component is shared by most of the models. For the time-averaged estimates of the natural variability, the ergodic variance has shown an increase of about 10% of the variability, which is due to the non-ergodic component of variability that acts synchronously between the members of a same model. Also, an ensemble has been reconstructed based on the SMP method by random sampling new members from a pool of climatic time periods. The resulting estimate of the natural variability has been shown to be smaller than both the time-averaged inter-member spread and the ergodic variance. We attribute this underestimation to an insufficiently large pool of time periods to choose from since some models have very few members.

The general effect of the weighting depends on both the structure of the ensemble and on the importance of the differences/consensuses between models' simulated natural variability. This characteristic should hence be considered in future ensembles. For example, the impact of imbalance in the sample sizes could become more important for future generations of models if higher levels of similarity appear between some of the models. Generally speaking, it is worth noting that unweighted variances are expected to be affected by larger sampling errors since giving a larger relative importance to the poorest estimates of the natural variability provided by the models with the smallest sample sizes (e.g. 2 members). In the present case, where such a weighting does not influence much the final multi-model estimate of the natural variability, the weighted coefficients should be preferred due to their smaller error. In an hypothetical case, where the weighting would have a significant impact, the use of an unweighted average between natural variability could be suitable by sustaining the democratic idea of "one model, one vote" in the construction of the multi-model estimate of the natural variability.

The unbalanced number of realisations in a multi-model ensemble necessarily affects the quantification of the other sources of uncertainty. An example that has been noted in Sect. 4.2.4 is that applying the ANOVA on an unbalanced dataset involves approxima-

tions in order to separate the variability into several components. More particularly, in order to assess the inter-model spread (not shown), one has to guess a constant sample size representing the entire ensemble. Consequently, if the sample sizes largely differ such as in CMIP3, this guess becomes rather uncertain and might have a large impact on the measure of the inter-model spread and the other sources of uncertainty (e.g. scenario).

The second source of uncertainty in climate-change projection emerges since several alternatives exist in order to build a climate model. While quantifying the inter-model spread might seem a relatively simple task, establishing a relationship with the so-called model uncertainty is a very complex one. This issue occurs mainly because models are not independent from one another and since there is no commonly accepted metric for quantifying model independence. In order to address this issue, we used the name of the research centres in order to focus on pairs of models that are likely to share important structural dependencies. In addition, other interesting pairs were those formed by rather distinct models but developed within a same institute, which may also share some dependencies in a more general sense. We proposed a theoretical framework in order to test the "same-institute assumption" that consists in attributing consensuses in the models' climate-change projections to both structural and institutional types of dependence.

A key point in the analysis resides in assessing the statistical significance of the difference between the climate-change signals from two models. The testing framework is based on a t-statistics of the difference of the means relatively to a measure of the natural variability. In a first attempt (Chap. 2), we used the inter-member spread for assessing the natural variability. However, such an approach might be questioned since very few members are generally available for each model. Moreover, some models providing a single realisation, some pairs of models with institutional dependence had to be excluded from the investigation.

Based on Chap. 3, it has been shown that single-model ensembles can be considered as ergodic, i.e. that the inter-member spread is assumed as equal to the time variability.

While this is expected for simulations run under stationary conditions, simulations with transient forcing can be also treated as if ergodic after detrending the time series with a 4^{th}-degree polynomial function, resulting in a slight increase in variability. The application of the ergodic assumption has been shown to be very important in the construction of a statistical test for assessing the significance of a difference between two models, especially when both are providing only a single realisation.

Among our findings, general consensuses have been found between models developed by a same institute. Striking examples consist in the pairs of models of CGCM and MIROC that differ by a change in resolution. Climate-change projections in summer surface air temperature were not statistically different over the land, recalling that these simulations have been interpolated over a coarser grid resolution. Also, successive versions of the GFDL model have shown very similar results over the entire North-American domain. A rather interesting result is the pair of models from UKMO, which are often considered as rather independent climate models. These have shown statistically similar results practically over the whole domain for the two first periods considered (2000-2020 and 2040-2060), while slight differences in their climate-change signal begin to be statistically different after about a century. The latter results should be interpreted carefully since it is also possible that both models lead to the same result in rather independent ways. In such a situation, further investigation should be spent on these two models in order to determine whether the consensus in the signal is informative or not.

Such consensuses between models developed by a same institute should at least be taken into account when considering a multi-model ensemble. Reducing the occurrence of potential non-informative consensuses in an ensemble allows a clearer interpretation of the inter-model spread as a measure of the modelling uncertainties, in addition to more direct relationship between consensuses and confidence in a specific climatic outcome. While excluding some models in an ensemble increases the error of the ensemble statistics, it could be understood as a cautious choice for reducing overconfidence. Since assessing model independence consists in a very complex task, this should be undertaken by the

entire climate modelling community. For example, an institute providing several models or versions of a same model to an inter-comparison project should at least provide some insights about the potential added value to the entire ensemble by considering several rather than a single model from their institute. From that point of view, it is also possible that two versions of a same model give a non-informative consensus in their response for a specific variable (e.g. temperature) while leading to an informative disagreement that contribute to the modelling uncertainty for another variable (e.g. precipitation).

In the case of large ensembles of opportunity such as the CMIP3 multi-model dataset, weak constraints are applied to the sampling procedure in order to favour a largest diversity of simulations over the structure of the ensemble. Forming ensembles in such an open way is likely to lead to unbalanced ensemble frameworks in the context of some specific studies. More precisely, an unbalanced sampling design leads to approximations in the statistical theory and is likely to lead to larger sampling errors compared to an optimised framework of the same size. Moreover, since the sampling of the models is done in a neither random or systematic manner, experts' judgement becomes necessary in order to interpret correctly the message conveyed by an ensemble of opportunity.

Appendix 4.A : Analysis of variance applied to a multi-model ensemble (MME)

In this section, we describe how the analysis of variance (ANOVA) can be applied to a multi-model ensemble in order to separate the total variability into two main components, i.e. the inter-model spread and the natural variability. We assume a MME formed by several models, each one providing an arbitrary number of realisations (at least two) of a given experiment (e.g. a given emission scenario).

Let us first describe one element of a multi-model ensemble of simulations according to the linear model

$$X_{mn} = \mu + a_m + e_{mn}, \tag{4.11}$$

where μ is the theoretical mean of the population, a_m the effect of the treatment (different models) and e_{mn} the residual error (natural variability) assumed as independent and identically distributed (*iid*) according to the model and member indices (m and n respectively). According to the present statistical model, the natural variability has to be assumed as the same across the models, what has been shown to be false in Chap. 3. We however assume by simplicity that e_{mn} is *iid* while some direct implications of this erroneous assumption will be discussed further.

As a starting point, we use the unweighted ensemble mean ($\overline{X}_{oo}$ in "o" notation),

$$\overline{X}_{oo} = \frac{1}{M}\sum_{m}^{M}\frac{1}{N_m}\sum_{n}^{N_m}X_{mn}, \tag{4.12}$$

an averaging that gives equal weighting to each of the models even when sample sizes differ. In order to study the components of variance, the total sum of square (SST) is written relatively to the unweighted ensemble mean as

$$SST = \sum_{m}^{M}\sum_{n}^{N_m}(X_{mn} - \overline{X}_{oo})^2 \tag{4.13}$$

where M is the number of models in the ensemble and N_m the sample size of the m^{th}

model. (4.13) can be decomposed as $SST = SSA + SSE$, where

$$SSA = \sum_{m}^{M} N_m(\overline{X}_{mo} - \overline{X}_{oo})^2 \tag{4.14}$$

and

$$SSE = \sum_{m}^{M}\sum_{n}^{N_m}(X_{mn} - \overline{X}_{mo})^2 \tag{4.15}$$

respectively. In order to interpret the sums (4.13) to (4.15) as variances, we have first to determine their respective number of degrees of freedom (df). In SST, N square terms are summed ($N = \sum_m N_m$) and the deviations used in the calculation are all independent with exception of one term. The latter is fixed by the $N-1$ other terms through the unweighted ensemble mean (4.12), and hence, (4.13) has $N-1$ df. Similarly, SSE contains N terms but M of them are constrained by the single-model ensemble means ($\overline{X}_{mo}$). It follows that the sum (4.15) has $N-M$ df. Finally, SSA consists in $M-1$ independent terms (or df) due to the multi-model ensemble average ($\overline{X}_{oo}$) that constrains the m^{th} single-model ensemble mean.

Dividing (4.15) by its number of df results in a mean squared error component (MSE), which can be interpreted as a measure of natural variability that is pooled across the models ($\hat{\sigma}^2_{WI}$), i.e.

$$\hat{\sigma}^2_{WI} = \frac{1}{N-M}\sum_{m}^{M}\sum_{n}^{N_m}(X_{mn} - \overline{X}_{mo})^2. \tag{4.16}$$

Under the *iid* assumption and with sufficiently large sample sizes, $\hat{\sigma}^2_{WI}$ should tend toward a theoretical value of natural variability that characterises a model and that is approached with sufficiently large sample size. However, since the *iid* assumption appears to be false (according to Chap. 3) and that the models generally provide different sample sizes (N_m), $\hat{\sigma}^2_{WI}$ is necessarily biased toward the models with the largest number of members.

Using the number of df is not sufficient in order to interpret the mean square of SSA as the inter-model variance. It can be seen by replacing (4.11) into (4.14) to obtain the

expected value of SSA as

$$E(SSA) = \sum_{m}^{M} N_m a_m^2 + \sum_{m}^{M} N_m (\overline{e}_{mo} - \overline{e}_{oo})^2. \tag{4.17}$$

While the inter-model variance ($\hat{\sigma}_M^2$) should be defined as $\frac{1}{M-1}\sum_m^M a_m^2$ according to (4.11), such a term can not be isolated analytically from (4.17) due to the unequal N_m. However, a strategy for circumventing this issue is to assume a constant sample size (N_0) that is representative of the ensemble in the expectation of SSA as :

$$E(SSA) \approx N_0 \sum_{m}^{M} a_m^2 + N_0 \sum_{m}^{M} (\overline{e}_{mo} - \overline{e}_{oo})^2. \tag{4.18}$$

It can be shown that the second term on the right-hand side of (4.18) corresponds to $(M-1)\sigma_{WI}^2$ and then that

$$\frac{E(SSA)}{M-1} = \sigma_{WI}^2 + N_0 \sigma_M^2. \tag{4.19}$$

It follows that the inter-model variance can be estimated as :

$$\hat{\sigma}_M^2 = \frac{SSA/(M-1) - SSE/(N-M)}{N_0}. \tag{4.20}$$

where the median of the sample sizes can be used as an educated guess for N_0.

Appendix 4.B : Assessing the natural variability by using the member-sampling method

In this section, we make an analytic demonstration of the uncertainty obtained using the member-sampling approach presented in Chap. 1. The resulting value of the uncertainty affecting the ensemble statistics will next be scaled in order to obtain an estimate of the multi-model natural variability. As shown previously in a perfect-ensemble experiment (Appendix 1.A), the member-sampling approach underestimates the expected effect of

the natural variability. This systematic bias will be discussed and related to the insufficient number of members available for each model in the ensemble. The imbalance in the sampling of the realisations across models will also be related to the obtained estimate of multi-model natural variability. For consistence of the demonstration, we assume the same basic constraint of selection as in the member sampling, i.e. one member per model is retained in each ensemble. By simplicity, we explicitly use an ensemble size of two models but a generalisation to an arbitrary number of models (M) will be provided through the demonstration.

Let X_{1j} and X_{2k} represent two models ($M = 2$) where both of the j and k indices refer to a particular member available for one of these models. The sample sizes for these models are denoted as N_1 and N_2 respectively. According to the member-sampling method, the multi-model mean is first calculated for each generated ensembles, which differ only by the retained member for each model, and hence the multi-model means are averaged. The latter overall ensemble mean ($\hat{\mu}$) can be written as follows :

$$\hat{\mu} = \frac{1}{N_1 N_2} \sum_{j}^{N_1} \sum_{k}^{N_2} \frac{X_{1j} + X_{2k}}{M}. \tag{4.21}$$

While the member-sampling method uses a random selection of members, we use for convenience in (4.21) all the possible combinations of members in a systematic manner. By rearranging the terms, we obtain

$$\hat{\mu} = \frac{1}{M} \left(\frac{.1}{N_1} \sum_{j}^{N_1} X_{1j} + \frac{1}{N_2} \sum_{k}^{N_2} X_{2k} \right). \tag{4.22}$$

The previous equation consists in the average of two single-model ensemble means and then can be generalised for an arbitrary number of models as

$$\hat{\mu} = \frac{1}{M} \sum_{m}^{M} \frac{1}{N_m} \sum_{n}^{N_m} X_{mn}, \tag{4.23}$$

i.e. the unweighted ensemble mean, which gives equal importance to the different models

no matter on how many members they provides. This is due to the fact that the average over the members is calculated before that over the models.

Let us now calculate the statistical error of the ensemble mean, using again two models as a starting point. According to the member-sampling method, the variance of the mean is calculated as

$$Var(\hat{\mu}) = \frac{1}{N_1 N_2 - 1} \sum_{j}^{N_1} \sum_{k}^{N_2} \left(\frac{X_{1j} + X_{2k}}{M} - \frac{\overline{X}_{1o} + \overline{X}_{2o}}{M} \right)^2 \tag{4.24}$$

where the first term in the parenthesis is the $(j,k)^{th}$ ensemble mean and the second the overall ensemble mean (Eq. 4.23). By rearranging the terms in (4.24), we obtain :

$$Var(\hat{\mu}) = \frac{1}{M^2 \times (N_1 N_2 - 1)} \sum_{j}^{N_1} \sum_{k}^{N_2} \left((X_{1j} - \overline{X}_{1o}) + (X_{2k} - \overline{X}_{2o}) \right)^2 . \tag{4.25}$$

By developing the square term in the sum, it can be shown that the cross product vanishes since the j and k indices vary independently. We hence obtain :

$$Var(\hat{\mu}) = \frac{1}{M^2} \left(\frac{1}{N_1 - \frac{1}{N_2}} \sum_{j}^{N_1} (X_{1j} - \overline{X}_{1o})^2 + \frac{1}{N_2 - \frac{1}{N_1}} \sum_{k}^{N_2} (X_{2k} - \overline{X}_{2o})^2 \right), \tag{4.26}$$

or in more general terms for an ensemble with an arbitrary number of models :

$$Var(\hat{\mu}) = \frac{1}{M^2} \sum_{m}^{M} \left(\frac{1}{N_m - \frac{1}{\prod_{m' \neq m} N_{m'}}} \sum_{n}^{N_m} (X_{mn} - \overline{X}_{mo})^2 \right) \tag{4.27}$$

For typical ensemble sizes, the fraction at the denominator is very small and can be assumed to be zero. Recalling that the standard error of the mean corresponds to the variance of the sample mean divided by the sample size, we obtain an estimate of the natural variability by multiplying (4.27) by M as

$$\hat{\sigma}^2_{mem} = \frac{1}{M} \sum_{m}^{M} \frac{1}{N_m} \sum_{j}^{N_m} (X_{mn} - \overline{X}_{mo})^2 \tag{4.28}$$

that consists in the average between M biased estimates of variance, each one representing the inter-member variability of a single model. Under the assumption that the N_m are sufficiently large, the bias of the estimates decreases asymptotically to zero. On the other hand, it is worth noting that this estimate of the natural variability is unweighted, i.e. that it is not biased toward particular models, for example those providing larger sample sizes.

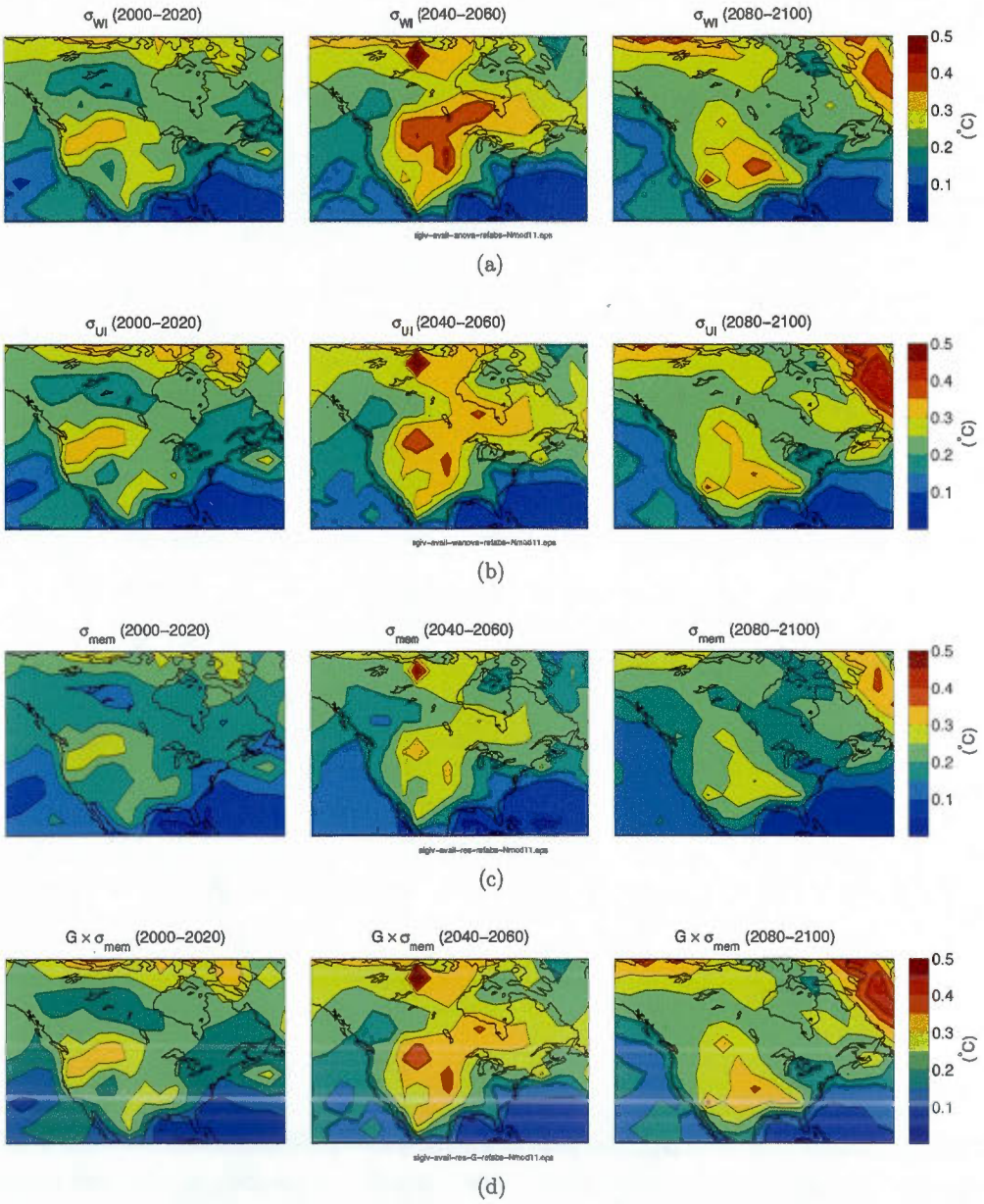
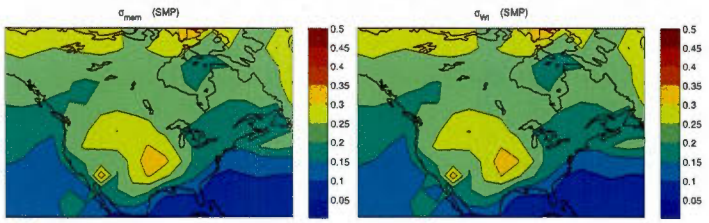
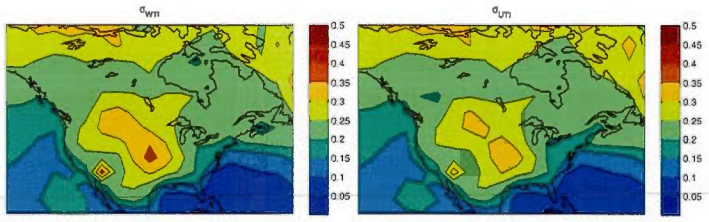


Fig. 4.1: Assessing the natural variability from a multi-model ensemble by using different estimators based on the inter-member spread : a) the weighted inter-member spread ($\hat{\sigma}_{WI}$; ANOVA mean-square error), b) the unweighted inter-member spread ($\hat{\sigma}_{UI}$), c) the member-sampling estimator ($\hat{\sigma}_{mem}$) and d) the empirically corrected member-sampling estimator ($G \times \hat{\sigma}_{mem}$).



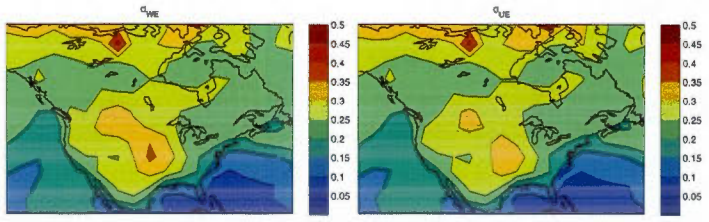
(a) (b)

Fig. 4.2: Assessing the natural variability from a multi-model ensemble where each single-model ensemble is reconstructed up to 100 members based on the single-model pooling (SMP) method : a) the member-sampling estimator ($\hat{\sigma}_{mem}$) and b) the ANOVA coefficient ($\hat{\sigma}_{WI}$).



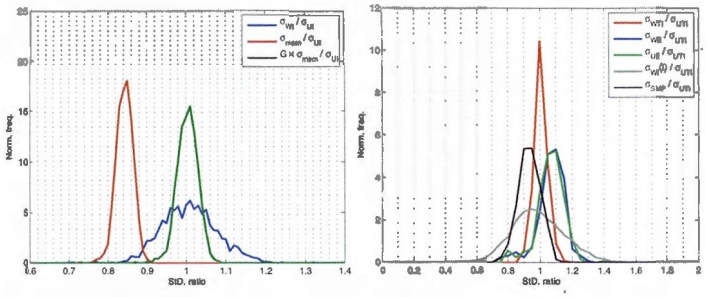
(a) (b)

Fig. 4.3: Assessing the natural variability from a multi-model ensemble by using a) weighted ($\hat{\sigma}_{WTI}$) and b) unweighted ($\hat{\sigma}_{UTI}$) time-averaged inter-member spreads.



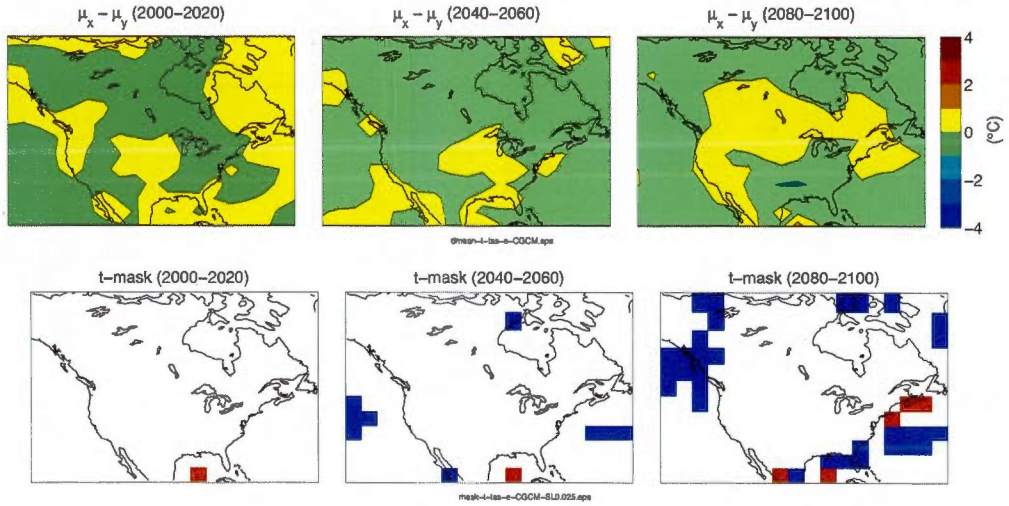
(a) (b)

Fig. 4.4: Assessing the natural variability from a multi-model ensemble by using a) weighted ($\hat{\sigma}_{WE}$) and b) unweighted ($\hat{\sigma}_{UE}$) ergodic variances.

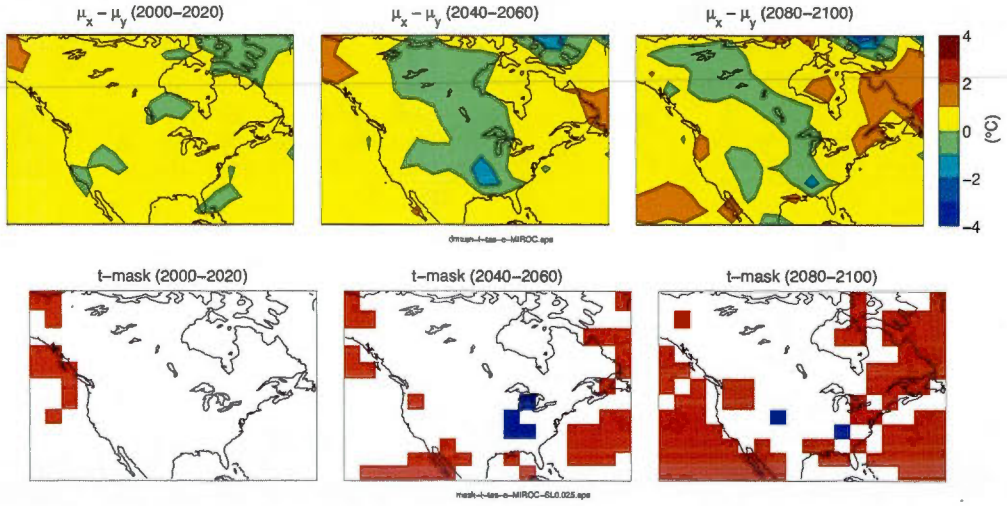


(a) (b)

Fig. 4.5: Ratio between the different estimates of the natural variability relatively to a reference estimator a) the unweighted inter-member spread ($\hat{\sigma}_{UI}$) and b) the unweighted time-averaged inter-member spread ($\hat{\sigma}_{UTI}$). Distributions are constructed using data from all grid points of the domain and all available 20-year average windows from 1900 to 2100.

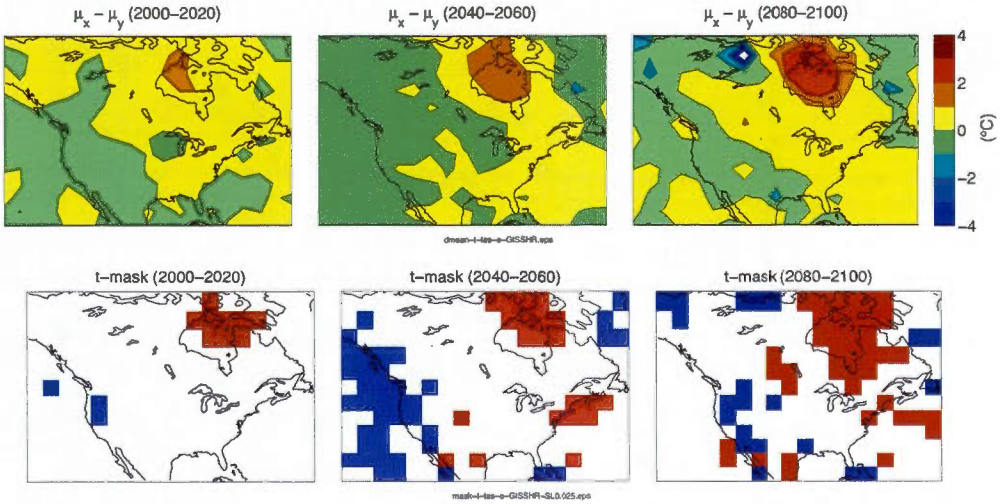


(a) CGCM : T47-T63

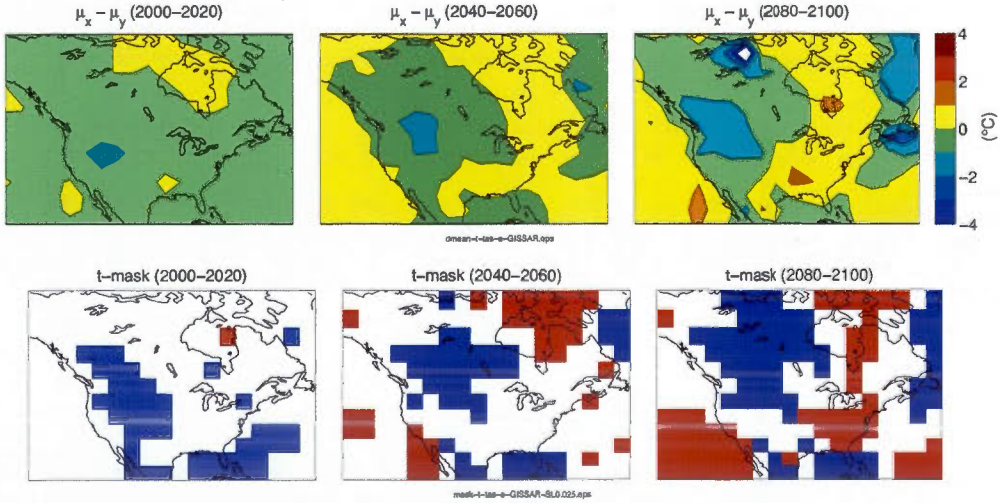


(b) MIROC : T106-T42

Fig. 4.6: To be continued...

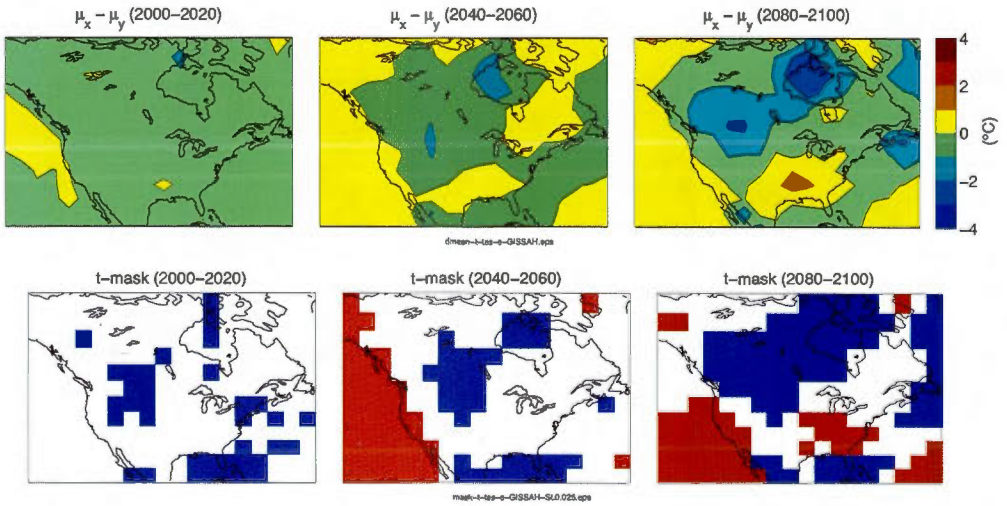


(c) GISS : EH-ER

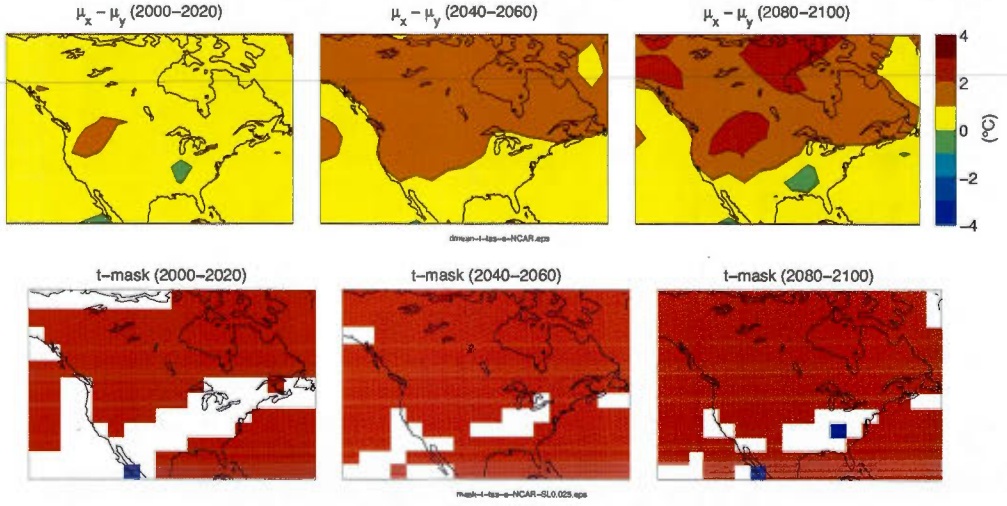


(d) GISS : AOM-ER

Fig. 4.6: To be continued...

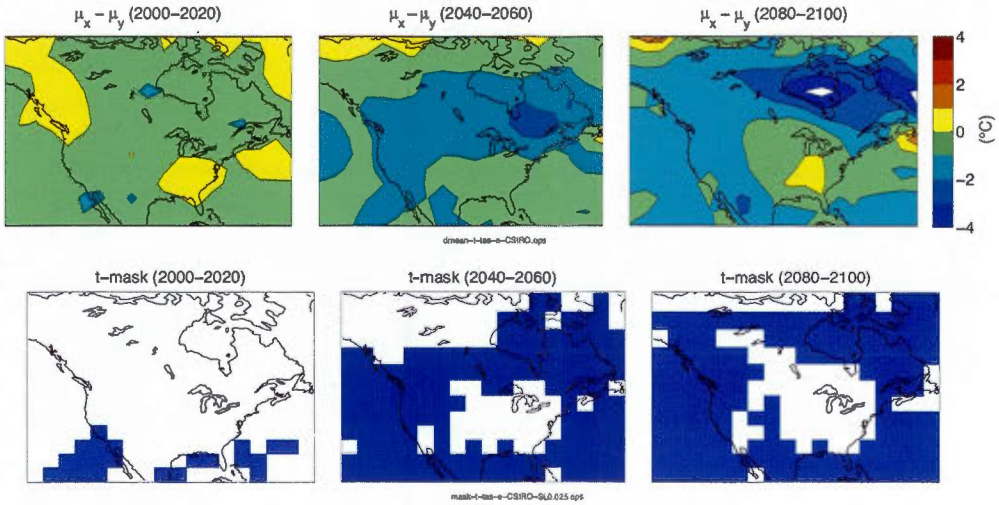


(e) GISS : AOM-EH

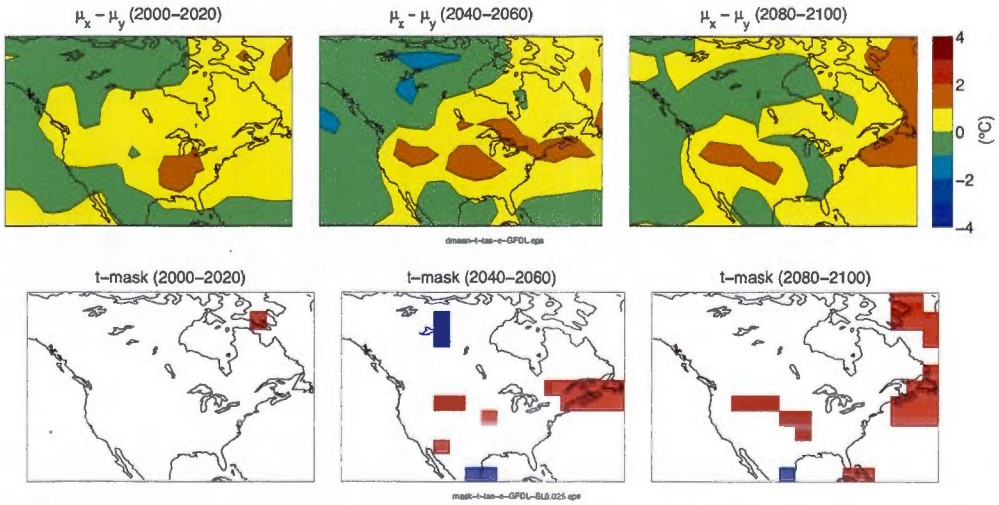


(f) NCAR : CCSM3-PCM

Fig. 4.6: To be continued...

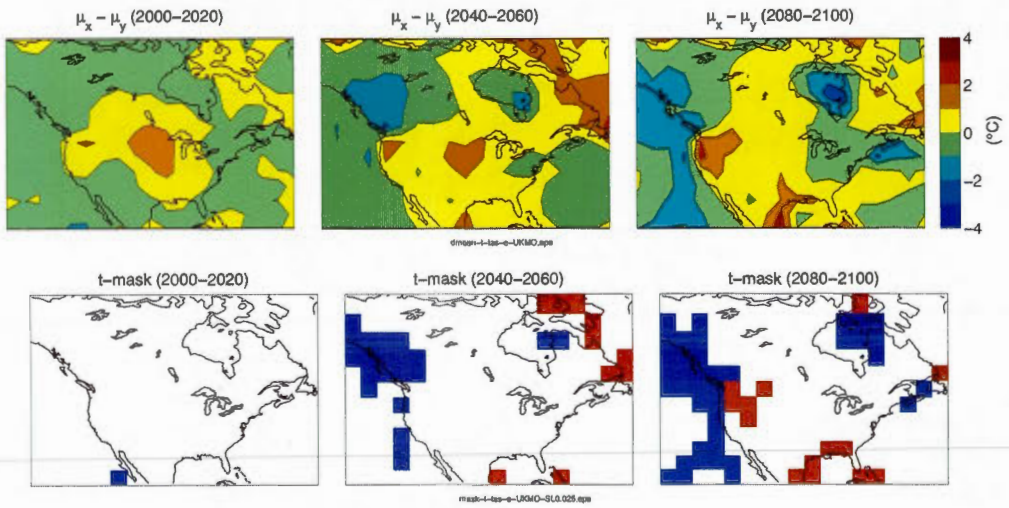


(g) CSIRO : 3.0-3.5



(h) GFDL : 2.0-2.1

Fig. 4.6: To be continued...



(i) UKMO : CM3-GEM1

Fig. 4.6: Difference of the ensemble mean climate-change signal for different pairs of models (or versions) developed by the same research institute. The climate-change signal is calculated for each simulation relatively to the 1980-2000 period. The panel at the bottom of each difference shows the mask of rejection of the null hypothesis by using a two-tailed t-test at the 5% significance level (2.5% on each side) based on the ergodic assumption. Red and blue colours mean positive and negative differences respectively.

CONCLUSION

Depuis une vingtaine d'années, une grande partie des connaissances scientifiques sur le système climatique a été acquise grâce à l'utilisation d'ensembles de simulations de type "multi-modèles". En guise d'exemples importants, les ensembles CMIP1 (Lambert et Boer, 2001) et CMIP3 (Meehl et al., 2007b) ont été respectivement les pierres d'assise des troisième et quatrième rapports du *Groupe d'experts intergouvernemental sur l'évolution du climat* (GIEC) (IPCC 2001, IPCC 2007) ; quant à CMIP5, il joue un rôle similaire dans le cadre du cinquième rapport. Ces étapes importantes témoignent de l'évolution des règles de l'art quant à notre compréhension du système climatique observé et de son évolution dans le futur, tout comme de la succession des différentes générations de modèles dont le niveau complexité est toujours grandissant. Le type d'ensemble à la base des rapports du GIEC est souvent appelé "ensemble d'opportunité" puisque basé sur la contribution de divers centres de recherches à l'échelle mondiale et dont le niveau de participation dépend des ressources et intérêts propres à chaque groupe.

La principale valeur de ces ensembles réside dans le fait qu'ils sont formés par des simulations provenant de modèles développés par différents centres de recherche et d'une manière, *a priori*, plutôt indépendante. Ces modèles consistant en différentes représentations mathématiques du système climatique, ils permettent un certain échantillonnage de l'incertitude scientifique autour de leur construction, c'est-à-dire quant aux hypothèses et approximations qui y sont employées. Ces ensembles sont riches en information, mais ils sont aussi extrêmement coûteux à produire et leur volume de données est imposant. Leur traitement peut poser des problèmes tant au niveau technique que pour des raisons plus conceptuelles, voire philosophiques. Cette thèse a permis de mettre en lumière plusieurs difficultés associées à l'utilisation d'ensembles multi-modèles de projections climatiques. Ce projet de recherche s'est positionné dans un cadre assez général de sorte

que la plupart des problématiques abordées puissent aussi s'appliquer à d'autres cas d'ensembles (e.g. CMIP5). L'ensemble CMIP3 a donc été utilisé à titre d'exemple d'ensemble d'opportunité. Cet ensemble étant assez vaste, l'analyse a été concentrée sur le champ de température de l'air à la surface à partir des simulations disponibles pour les 20^e et 21^e siècles, où des forçages externes sont appliqués comme les émissions de gaz à effet de serre et d'aérosols (GESA). De plus, le cadre de la recherche s'est restreint aux deux sources d'incertitudes fondamentales dans les projections climatiques, soient la variabilité climatique naturelle et l'incertitude "modèle". Le scénario d'émissions A1B a été choisi puisqu'il contient le plus grand nombre de modèles et de simulations.

L'échantillonnage d'un ensemble d'opportunité étant principalement basé sur la participation des divers centres de recherches, et donc d'une certaine manière de l'offre en simulations, la structure finale d'un tel ensemble est définie par des facteurs externes plutôt que par un cadre expérimental précis permettant une optimisation de l'analyse subséquente (c.-à-d. une décomposition des différentes composantes d'incertitude). En conséquence, si on interprète l'ensemble CMIP3 comme une matrice de simulations dont un axe représente les modèles et l'autre les membres, il apparaît que certains éléments sont manquants en raison de l'échantillonnage des membres qui varie substantiellement entre les modèles. Une question à la base du premier chapitre de la thèse était d'estimer la perte en information reliée à ce manque apparent de membres dans l'ensemble. La matrice incomplète étant la seule de disponible, cette question n'a pu être abordée directement. Nous avons donc procédé à la quantification de la perte d'information intervenant lors d'un second processus d'échantillonnage depuis l'ensemble original.

Ce second processus d'échantillonnage est inspiré de ce qui se fait généralement par les experts lorsqu'un ensemble est jugée trop volumineux pour être traité en entier. Par exemple, un centre de modélisation régionale du climat ne peut généralement pas traiter toutes les simulations provenant de tous les modèles de CMIP3, et donc une présélection de simulations s'avère alors nécessaire. Une décision typique visant à réduire la taille d'un tel ensemble consiste à n'utiliser qu'un seul membre par modèle (p. ex. Bombardi

et Carvalho, 2011 ; Peings et Douville, 2010 ; Räisänen et al. 2010). Ce choix se justifie généralement sur le fait que l'incertitude modèle est plutôt grande en comparaison avec la variabilité climatique naturelle (Hawkins et Sutton, 2011). Ce processus de sélection a été utilisé comme contrainte de base dans l'élaboration d'un cadre expérimental visant à mesurer l'effet d'une présélection des simulations.

À partir de l'ensemble CMIP3, la sélection d'un sous-ensembles de la forme "un membre par modèle" laisse plus d'un million de possibilités. En utilisant une approche par *bootstrap*, de tels ensembles ont pu être échantillonnés en grand nombre. Pour chaque ensemble ainsi formé, deux statistiques d'ensemble ont été calculées, soient la moyenne d'ensemble et l'écart-type inter-modèle. En considérant plusieurs milliers d'ensembles, des statistiques ont pu être calculées sur ces statistiques d'ensemble. En particulier, l'écart-type des statistiques d'ensemble a été interprétée comme une mesure d'incertitude reliée à la présélection d'un ensemble. Via la sélection aléatoire d'un membre par modèle, cette incertitude est une manifestation de la variabilité inter-membre, et donc de la variabilité climatique naturelle simulée par les modèles. Un résultat intéressant est que cette mesure d'incertitude sous-estime l'effet qui serait attendu si l'ensemble de départ avait compté un très grand nombre de membres par modèle. Ceci a été démontré à l'aide d'une expérience de type Monte-Carlo où un ensemble parfait était comparé au cas imparfait dont la structure est identique à l'ensemble CMIP3 disponible. Du même coup, cette expérience a permis de calculer un facteur empirique permettant de corriger l'ampleur de l'incertitude attribuable au choix des membres pour qu'elle corresponde à la valeur attendue pour un ensemble contenant un grand nombre de membres. Ceci met en lumière une différence importante entre l'incertitude perçue lors de la sélection et sa valeur attendue par l'effet réel de la variabilité climatique naturelle.

De manière similaire, le processus de présélection d'un groupe de modèles a aussi été étudié dans le premier chapitre. Une particularité de la méthode utilisée est l'utilisation de l'hypothèse voulant que les modèles de CMIP3 consistent en un échantillon représentatif d'une plus grande population. Par exemple, on pourrait imaginer une telle population

comme étant formée de modèles hypothétiques dont le niveau de complexité serait similaire à ceux formant l'ensemble CMIP3. Techniquement, cette hypothèse revient à permettre le remplacement des modèles lors de l'échantillonnage par *bootstrap*. Cette généralisation de la méthode d'échantillonnage des modèles a été adoptée devant l'approche généralement utilisée de contraindre la sélection à des modèles différents. Ceci met en lumière une fois de plus la différence entre l'incertitude perçue par différentes sélections d'ensembles de l'incertitude réelle qui serait attendue à l'aide d'un ensemble contenant cette fois un très grand nombre de modèles. Un cas particulier consiste en un ensemble de 24 modèles où un seul choix de sous-ensemble peut être fait sous la contrainte d'utilisation de modèles différents, tandis que la généralisation par remplacement des modèles mène à $1,6 \times 10^{13}$ ensembles différant par au moins un modèle. Cette approche donne notamment une mesure d'incertitude reliée directement à la taille de l'échantillon et à l'écart-type inter-modèle, soit la relation de l'erreur type de la moyenne (von Storch et Zwiers, 1999).

L'incertitude due à la sélection des modèles s'est montrée généralement plus grande que celle associée à la sélection des membres. Tel qu'attendu par la loi de l'erreur type, les deux sources d'incertitudes deviennent de plus en plus importantes au fur et à mesure que la taille de l'ensemble est réduite. L'incertitude reliée à la sélection des membres étant à peu près constante dans le temps, son importance diminue relativement au signal de changement climatique lorsqu'on avance dans le futur. Pour l'incertitude associée à la sélection des modèles, celle-ci augmente avec le temps puisque directement reliée à l'écart-type inter-modèle du signal de changement climatique. Par ailleurs, l'incertitude associée à la sélection des modèles tend à rester constante dans le temps par rapport au signal de changement climatique, un résultat attribué à la chance puisque l'écart-type inter-modèle dépend des différences structurelles entre modèles ainsi que de leurs différentes réponses au forçage externe, tandis que le signal de changement climatique dépend de l'intensité des émissions de GESA.

Le choix d'un membre par modèle étant généralement fait aléatoirement en pratique,

le cadre expérimental proposé est très représentatif de ce qui se fait dans la réalité. La sélection d'un membre par modèle apparaît comme une technique très efficace pour réduire la taille d'un ensemble volumineux puisque son effet sur les statistiques d'ensemble est assez petit, même pour des ensembles comptant peu de modèles. D'un autre côté, le cadre proposé pour la sélection des modèles est peu contraint comparativement à une présélection par les experts, où des processus complexes de sélection de modèles peuvent être considérés. Le cadre expérimental proposé peut permettre de juger si l'effet du choix des modèles sur les statistiques d'ensemble est suffisamment grand pour s'engager dans un tel processus et y allouer beaucoup de ressources. Basé sur des contraintes bien définies, il semble qu'effectuer une présélection de modèles peut en principe permettre de corriger certaines lacunes apparaissant dans l'échantillonnage initial d'un ensemble d'opportunité. Par exemple, les notions de performance (Giorgi et Mearns, 2002) ou d'indépendance (Whetton et al. 2007; Abramowitz et Gupta, 2008) des modèles peuvent permettre d'améliorer la qualité de l'échantillon initial, mais encore faut-il que les contraintes de sélection soient robustes et fassent consensus entre les scientifiques.

Après avoir étudié les caractéristiques d'un échantillonnage de type "expert" dans le premier chapitre, le second chapitre s'est concentré sur le processus d'échantillonnage à la base même de la formation d'un ensemble d'opportunité. Dans un tel ensemble, l'échantillonnage des modèles se fait d'une manière ni aléatoire ni systématique, ce qui engendre plusieurs difficultés au niveau de l'interprétation des résultats de l'ensemble. Bien que les modèles soient construits différemment, certaines particularités leurs sont souvent communes, comme certaines hypothèses sur les processus physiques d'intérêt à y intégrer ou quant à la manière de les transposer sous forme d'équations. La présence de similarités au niveau de la structure des modèles est en quelque sorte attendue étant donné la manière dont la science évolue. Par exemple, les experts partagent des connaissances sur le système climatique et quant à la manière de construire les modèles. Ce manque d'indépendance apparent entre les modèles étant donc compréhensible, il n'en est pas moins que la portée du problème est assez mal comprise. Une raison à ceci est qu'aucune métrique permettant d'évaluer l'indépendance des modèles ne fait présente-

ment consensus entre les scientifiques (Tebaldi et Knutti, 2007). Ce problème est un obstacle majeur devant toute interprétation probabiliste des résultats de l'ensemble. En particulier, on s'attendrait qu'un échantillon de modèle totalement indépendants mène aux propriétés suivantes :

1. Un consensus entre plusieurs modèles est un indicateur de confiance en un résultat donné
2. L'écart-type inter-modèle peut être interprété comme une mesure de l'incertitude modèle
3. L'erreur de la moyenne par rapport au climat réel devrait diminuer au fur et à mesure que des modèles sont ajoutés à l'ensemble.

Il a été montré que certains biais par rapport aux observations sont partiellement corrélés entre les modèles (e.g. Lambert et Boer, 2001 ; Knutti et al. 2010), ce qui serait l'indicateur d'un manque d'indépendance. Cependant, ce type d'approche nécessite des données d'observation et donc ne peut être directement appliquée au cas des changements climatiques attendus pour le prochain siècle, à moins d'utiliser l'hypothèse que deux représentations indépendantes du climat observé sont aussi des représentations indépendantes du système climatique quant à sa sensibilité aux forçages de GESA. Une autre approche consiste à étudier les similarités entre les sorties de modèles sans avoir recours aux observations. Par exemple, Masson et Knutti (2011) ont montré que les modèles développés par une même institution tendent à mener vers des résultats similaires.

En guise d'approche au problème de l'indépendance, nous avons choisi d'étudier la nature des consensus de changements climatiques entre modèles développés par une même institution. Dans ce contexte, les modèles de CMIP3 qui sont développés par un même groupe de recherche sont souvent très similaires au niveau de leur construction. Certaines paires de modèles diffèrent seulement par la résolution (paires CGCM et MIROC), ou consistent en des versions successives d'un même modèle (paires CSIRO et GFDL) où des changements relativement mineurs sont apportés au code des modèles. D'autres modèles

développés par une même institution sont caractérisés par des différences structurelles à plus haut niveau, par exemple les trois modèles GISS diffèrent par leur composante d'atmosphère, d'océan, de surface terrestre, de glace de mer et de couplage. Finalement, les paires NCAR et UKMO comprennent chacunes deux modèles assez différents mais qui ont tout de même été développés par une même institution.

Le cadre expérimental proposé vise à établir un critère permettant d'invalider un consensus entre deux modèles lorsque des raisons suffisantes peuvent remettre en cause leur indépendance. D'abord, afin de déterminer s'il y a consensus, la différence entre les signaux de changements climatiques a été comparée à la variabilité climatique naturelle simulée par les modèles. Les différences qui sont statistiquement non-significatives ont été interprétées comme des consensus entre modèles. Bien qu'il soit possible qu'un tel consensus apparaisse aussi entre deux modèles indépendants, le lien structurel ou institutionnel entre les modèles faisant consensus pourrait apparaître comme une raison suffisante pour remettre en cause leur indépendance, et donc de rejeter leur consensus. En pratique, le rejet d'un consensus pourrait consister à ne considérer qu'un seul modèle de la paire lors du calcul du signal de changement climatique moyen dans l'ensemble CMIP3. Un tel filtrage de l'ensemble permettrait de clarifier la notion d'un consensus dans l'ensemble (p. ex. via la moyenne) qui devrait apparaître par annulation des erreurs plutôt que par une corrélation de celles-ci. Du même coup, le filtrage d'un ensemble permettrait de clarifier la relation entre l'écart-type inter-modèle et la notion d'incertitude modèle.

Les résultats de l'analyse par paire de modèles développés par une même institution ont révélé plusieurs consensus. Les paires CGCM et MIROC ont montrés des signaux de changement climatiques très similaires. Il faut dire que les résultats de ces modèles ont été interpolés sur une grille commune, et donc qu'une partie importante de la valeur potentiellement ajoutée par le modèle à haute résolution (Di Luca et al., 2013) n'était pas considérée. La comparaison entre les modèles GISS munis de la même composante atmosphérique (GISS-EH et ER) a quant à elle donné de forts consensus principalement

au-dessus du continent. La paire GFDL a aussi mené à un consensus sur la majeure partie du domaine, suggérant que les changements structurels entre ces deux versions du même modèle étaient relativement mineurs. Un autre consensus a été trouvé entre les modèles UKMO. Ces derniers sont apparemment très différents d'un point de vue structurel mais partagent tout de même des similarités au niveau du schéma radiatif.

En principe, l'indépendance des modèles de climat devrait pouvoir être évaluée via une comparaison exhaustive du code de chacun des modèles, avec une attention particulière aux hypothèses et approximations utilisées. Il est assez évident que cette avenue est extrêmement laborieuse. Dans ce sens, la problématique de l'indépendance des modèles devrait être prise au sérieux par l'ensemble de la communauté des sciences du climat. Par exemple, un centre fournissant à un ensemble des simulations provenant de deux de ses modèles, ou versions d'un même modèle, devrait fournir une certaine justification quant à la valeur ajoutée lorsque sont considérés les deux modèles plutôt qu'un seul. Ceci pourrait grandement aider à l'interprétation de l'ensemble final et ses utilisateurs pourraient être en quelque sorte orientés vers la sélection d'un groupe de modèles plutôt indépendants. Poussant cette idée un peu plus loin, au lieu de traiter le problème *a posteriori* par un filtrage de l'ensemble, on pourrait aussi imaginer l'ajout de telles contraintes lors du processus d'échantillonnage de l'ensemble. Ainsi, un groupe de modélisation ne pouvant justifier la valeur ajoutée par un second modèle devrait choisir elle-même la version à soumettre à l'analyse par la communauté. Le filtrage se ferait donc par les développeurs qui connaissent bien ces modèles plutôt que par la communauté qui doit souvent se limiter à une interprétation de la documentation fournie. Ce type de contrainte ajoutée au processus d'échantillonnage pourrait en plus offrir certaines possibilités quant à une redistribution des ressources informatiques, par exemple en produisant plus de membres pour le modèle choisi ou même en produisant des simulations supplémentaires à l'aide d'un modèle développé par un centre bénéficiant de moins grandes ressources.

Dans le cadre de ce travail, certaines difficultés dans l'analyse ont été reliées au pauvre échantillonnage de membres pour certains des modèles. Par exemple, dans le chapitre

1, les modèles avec le plus petit nombre de membres ont été filtrés de l'ensemble en vue de minimiser le biais systématique de l'incertitude reliée au choix d'un membre par modèle. Au cours du deuxième chapitre, le fait que certains modèles ne fournissait qu'un seul membre a nécessité le rejet des modèles développées par certains centres, ces derniers ne pouvant être comparés à l'aide d'un t-test basé sur la variabilité inter-membre. Un objectif visé dans le troisième chapitre était de s'attaquer à ce problème en "remplissant" les éléments considérés comme manquants dans la matrice de simulations par des méthodes peu coûteuses relativement à l'utilisation d'un modèle de circulation générale couplé.

Au cours du chapitre 3, un cadre décisionnel a été défini en vue de choisir l'approche la mieux adaptée à l'ensemble utilisé. Ce cadre proposait deux approches, soit 1) utiliser l'information temporelle d'un modèle pour lui générer des membres supplémentaires, ou 2) utiliser l'information temporelle de tous les modèles pour la reconstruction de membres. Le choix de l'approche s'est basé sur deux questions fondamentales reliées à la simulation de la variabilité climatique naturelle dans l'ensemble. La première consiste à évaluer si un ensemble de membres générés par un même modèle peut être considéré comme étant *ergodique*, c'est-à-dire que la variabilité entre les membres est à peu près égale à celle mesurée dans le temps. Bien que l'ergodicité soit attendue pour des simulations sans forçages externes (Peixoto et Oort, 1992), l'approche utilisée pour des simulations sous GESA fut d'abord de les rendre "stationnaires par traitement", soit en soustrayant les tendances (représentées par des polynômes) de leurs séries temporelles. La seconde question visait quant à elle à vérifier si la variabilité naturelle était égale entre les modèles. Selon le cadre décisionnel proposé, une réponse positive à la première question permettrait l'utilisation de l'approche 1, tandis qu'une réponse positive aux deux questions permettrait l'utilisation de l'approche 2. Pour les modèles considérés, l'hypothèse d'ergodicité s'est avérée plutôt vraie, tandis que l'hypothèse selon laquelle les modèles simulent le climat avec la même varaibilité naturelle s'est avérée plutôt fausse.

Le chapitre 4 a permis d'effectuer une synthèse des principaux concepts avancés dans cette thèse tout en proposant deux exemples d'applications concrètes. Le but visé par la première application était de comparer différentes approches afin d'estimer la variabilité climatique naturelle par une combinaison de l'information disponible provenant de tous les modèles de l'ensemble. Plusieurs estimateurs y ont été discutés. D'abord, une forme analytique de la variabilité inter-membre estimée à partir de la méthode de sélection des membres (Chap. 1) a été fournie. Cette démonstration mathématique permet d'abord de comprendre que le biais systématique de l'estimateur apparaît puisqu'il consiste en une moyenne multi-modèle sur plusieurs estimateurs biaisés de la variance inter-membre. De plus, il apparaît de cette démonstration que la méthode de sélection des membres mène à un estimé qui est non-pondéré, c'est-à-dire qui donne le même poids à chacun des modèles peu importe la taille des échantillons. Comme deuxième estimateur, la variabilité inter-membre a été calculée de manière plutôt classique en fonction des déviations autour de la moyenne d'ensemble (sur les membres) de chaque modèle. Cet estimateur provient de la méthode d'analyse de variance (ANOVA), et suppose que tous les modèles ont une variabilité inter-membre égale, ce qui n'est visiblement pas le cas. L'estimateur résulte donc en une variance pondérée selon la taille de l'échantillon de chaque modèle. Nous avons donc proposé une version non-pondérée de cet estimateur qui consiste à faire la moyenne multi-modèle des variances inter-membre (non-biaisées). Ensuite, ces deux derniers estimateurs ont été généralisés en considérant une moyenne au cours du temps, basée sur l'hypothèse que la variabilité inter-membre reste à peu près constante, ses variations étant principalement dues à de petits échantillons de membres. La variabilité inter-membre moyennée dans le temps a aussi été formulée sous des formes pondérées et non-pondérées par rapport à la taille des échantillons. Finalement, deux autres estimateurs ont été fournis, toujours en versions pondérées et non-pondérées, mais cette fois en supposant l'ergodicité dans les ensembles de membres. On peut voir ces estimateurs comme étant des variances temporelles moyennées sur les membres et les modèles, plutôt que des variances inter-membres moyennées sur le temps et les modèles comme ce fut le cas pour les estimateurs précédents.

La comparaison des différents estimateurs permettant de combiner les variabilités naturelles au niveau multi-modèle a généralement montré de faibles différences entre les versions pondérées et non-pondérées. La structure de l'ensemble utilisé est en partie responsable de ce résultat ainsi que la nature des différences entre les variabilités naturelles simulées par les modèles. D'une manière plus générale, la qualité des estimés pondérés devrait être supérieure à leur version non-pondérée respective, cette dernière donnant une plus grande importance relative aux mauvais estimés fournis par les modèles avec moins de membres. Cependant, dans un cas où les différences dues à la pondération seraient grande (p. ex. autre variable ou ensemble), les versions non-pondérées se doivent d'être considérées basé sur le principe démocratique de "un modèle un vote".

Une application simple du principe d'ergodicité a aussi permis de reconstruire des membres supplémentaires afin de "remplir" la matrice de simulations, tel que suggéré au chapitre 3. Cependant, l'application de la méthode de reconstruction était basée sur un échantillonnage aléatoire de périodes temporelles pour générer de nouveaux membres, ce qui a résulté en une sous-estimation de la variabilité naturelle en comparaison avec les autres estimateurs. De plus longues séries temporelles, et donc un plus grand choix de périodes temporelles, aurait probablement permis d'obtenir un estimé de la variabilité naturelle se rapprochant des autres estimateurs. D'un autre côté, une application systématique de l'hypothèse d'ergodicité, c'est-à-dire en moyennant la variabilité temporelle sur plusieurs membres, a montré une légère surestimation de la variabilité climatique naturelle par rapport aux autres méthodes. Cette variabilité supplémentaire est principalement due à la composante non-ergodique associée aux forçage résiduels (comme les émissions volcaniques) après soustraction des tendances dans les séries temporelles.

Dans le chapitre 4, le deuxième exemple d'application visait à améliorer les tests statistiques proposés au chapitre 2 en vue d'étendre la comparaison aux paires de modèles ayant été rejetées de l'analyse étant donné que chacun des modèles ne comportait qu'un seul membre. En utilisant l'hypothèse d'ergodicité, l'information temporelle de ces simulations a permis d'évaluer le niveau de signifiance statistique des différences entre

signaux de changements climatiques, et ce, même pour des modèles n'ayant qu'un seul membre. Entre autres, ce type d'approche aura permis de démontrer que l'information temporelle des simulations peut servir efficacement à palier le manque de membres de certains modèles dans l'ensemble. La notion d'ergodicité dans les ensembles peut donc mener à un certain questionnement quant au nombre de membres fournis dans un ensemble multi-modèle ainsi que de la valeur ajoutée par de longues simulations de l'ère préindustrielle. L'attribution de tendances basées sur des polynômes aux séries temporelles pouvant dépendre de la variable considérée et donc se devant d'être étudiée plus en profondeur, la possibilité d'obtenir un ensemble ergodique même en présence de forçages externes (GESA) s'avère une méthode assez efficace pour réduire le volume de données à extraire d'un grand ensemble. Par exemple, cette méthode pourrait être utilisée comme une alternative peu coûteuse comparativement au téléchargement des simulations stationnaires de la période préindustrielle dans l'ensemble CMIP3, qui consistent en des échantillonnages explicites de la variabilité naturelle simulée par les modèles.

REFERENCES

Abramowitz, G. (2010). Model independence in multi-model ensemble prediction. *Australian Meteorological and Oceanographic Journal*, 59 :3–6.

Abramowitz, G. and Gupta, H. (2008). Toward a model space and model independence metric. *Geophys. Res. Lett.*, 35(5) :L05705.

Alexandru, A., de Elia, R., and Laprise, R. (2007). Internal variability in regional climate downscaling at the seasonal scale. *Mon. Wea. Rev.*, 135 :3221–3238.

Allen, M. R. and Ingram, W. J. (2002). Constraints on future changes in climate and the hydrologic cycle. *Nature*, 419(6903) :224–232. 10.1038/nature01092.

Ammann, C. M., Meehl, G. A., Washington, W. M., and Zender, C. S. (2003). A monthly and latitudinally varying volcanic forcing dataset in simulations of 20th century climate. *Geophys. Res. Lett.*, 30(12) :1657.

Annan, J. D. and Hargreaves, J. C. (2010). Reliability of the CMIP3 ensemble. *Geophys. Res. Lett.*, 37(2) :L02703.

Annan, J. D. and Hargreaves, J. C. (2011). Understanding the CMIP3 multimodel ensemble. *Journal of Climate*, 24(16) :4529–4538.

Bechtold, P., Bazile, E., Guichard, F., Mascart, P., and Richard, E. (2001). A mass flux convection scheme for regional and global models. *Quart. J. Roy. Meteorol. Soc.*, 127 :869–886.

Bleck, R. (2002). An oceanic general circulation model framed in hybrid isopycnic-cartesian coordinates. *Ocean Modelling*, 4(1) :55–88.

Bombardi, R. and Carvalho, L. (2011). The south atlantic dipole and variations in the characteristics of the South American Monsoon in the WCRP-CMIP3 multi-model simulations. *Climate Dynamics*, (11-12) :2091–2102.

Bona, M. (2006). *A Walk Through Combinatorics : An Introduction to Enumeration and Graph Theory (Second Edition)*. World Scientific Publishing Company.

Christensen, J., Hewitson, B., Busuioc, A., Chen, A., Gao, X., Held, I., Jones, R., Kolli, R., Kwon, W.-T., Laprise, R., Magaña Rueda, V., Mearns, L., Menéndez, C., Räisänen, J., Rinke, A., Sarr, A., and Whetton, P. (2007). Regional Climate Projections. In : *Climate Change 2007 : The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Christensen, J. H., Kjellström, E., Giorgi, F., Lenderink, G., and Rummukainen, M. (2010). Weight assignment in regional climate models. *Climate Research*, 44(2-3) :179–194. 10.3354/cr00916.

Déqué, M., Rowell, D., Lüthi, D., Giorgi, F., Christensen, J., Rockel, B., Jacob, D., Kjellström, E., de Castro, M., and van den Hurk, B. (2007). An intercomparison of regional climate simulations for europe : assessing uncertainties in model projections. *Climatic Change*, 81(0) :53–70.

Déqué, M., Somot, S., Sanchez-Gomez, E., Goodess, C., Jacob, D., Lenderink, G., and Christensen, O. (2012). The spread amongst ensembles regional scenarios : regional climate models, driving general circulation models and interannual variability. *Climate Dynamics*, 38(5) :951–964.

Deser, C., Phillips, A., Bourdette, V., and Teng, H. (2010). Uncertainty in climate change projections : the role of internal variability. *Climate Dynamics*, 38(3-4) :527–546.

Di Luca, A., Elía, R., and Laprise, R. (2013). Potential for small scale added value of RCM's downscaled climate change signal. *Climate Dynamics*, 40(3-4) :601–618.

Dibike, Y. B., Gachon, P., St-Hilaire, A., Ouarda, T. B. M. J., and Nguyen, V. T. V. (2008). Uncertainty analysis of statistically downscaled temperature and precipitation regimes in northern canada. *Theoretical and Applied Climatology*, 91(1-4) :149–170.

Foley, A. (2010). Uncertainty in regional climate modelling : A review. *Progress in Physical Geography*, 34(5) :647–670.

Giorgi, F. and Mearns, L. O. (2002). Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) Method". *Journal of Climate*, 15(10) :1141–1158.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C. (2008). Performance metrics for climate models. *J. Geophys. Res.*, 113(D6) :1984–2012.

Greene, A. M., Goddard, L., and Lall, U. (2006). Probabilistic multimodel regional temperature change projections. *Journal of Climate*, 19(17) :4326–4343.

Hawkins, E. and Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, 90(8) :1095–1107.

Hawkins, E. and Sutton, R. (2011). The potential to narrow uncertainty in projections of regional precipitation change. *Climate Dynamics*, 37(1-2) :407–418.

Houle, D., Bouffard, A., Duchesne, L., Logan, T., and Harvey, R. (2012). Projections of future soil temperature and water content for three southern quebec forested sites. *Journal of Climate*, 25(21) :7690–7701.

IPCC (2001). *Climate Change 2001 :The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change* [Houghton, J.T., Y. Ding, D.J. Griggs, M. Noguer, P.J. van der Linden, X.

Dai, K. Maskell, and C.A. Johnson (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 881pp.

IPCC (2007). *Climate Change 2007 : The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 996 pp.

Jun, M., Knutti, R., and Nychka, D. W. (2008a). Local eigenvalue analysis of CMIP3 climate model errors. *Tellus A ; Vol 60, No 5 (2008).*

Jun, M., Knutti, R., and Nychka, D. W. (2008b). Spatial analysis to quantify numerical model bias and dependence. *Journal of the American Statistical Association*, 103(483) :934–947.

Kain, J. S. and Fritsch, J. M. (1990). A one-dimensional entraining/detraining plume model and application in convective parameterization. *J. Atmos. Sci.*, 47 :2784–2802.

Kendall, M. (1946). *The advanced theory of statistics. Vol. 2.* Griffin, London.

Knutti, R. (2010). The end of model democracy? *Climatic Change*, 102(3-4) :395–404.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A. (2010). Challenges in combining projections from multiple climate models. *Journal of Climate*, 23(10) :2739–2758.

Lambert, S. J. and Boer, G. J. (2001). CMIP1 evaluation and intercomparison of coupled climate models. *Climate Dynamics*, 17(2) :83–106.

Lean, J., Beer, J., and Bradley, R. (1995). Reconstruction of solar irradiance since 1610 : Implications for climate change. *Geophys. Res. Lett.*, 22(23) :3195–3198.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, 54 :421–431.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2) :130–141.

Lucas-Picher, P., Caya, D., de Élía, R., and Laprise, R. (2008). Investigation of regional climate models internal variability with a ten-member ensemble of 10-year simulations over a large domain. *Clim. Dyn.*, 31(7-8) :927–940.

Masson, D. and Knutti, R. (2011). Climate model genealogy. *Geophys. Res. Lett.*, 38(8) :L08703.

Mearns, L. O., Gutowski, W., Jones, R., Leung, R., McGinnis, S., Nunes, A., and Qian, Y. (2009). A regional climate change assessment program for North America. *Eos Trans. AGU*, 90(36).

Meehl, G., Stocker, T., Collins, W., Friedlingstein, P., Gaye, A., Gregory, J., Kitoh, A., Knutti, R., Murphy, J., Noda, A., Raper, S., Watterson, I., Weaver, A., and Zhao, Z.-C. (2007a). Global Climate Projections. In : *Climate Change 2007 : The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., and Stouffer, R. J. (2000). The Coupled Model Intercomparison Project (CMIP). *Bulletin of the American Meteorological Society*, 81(2) :313–318.

Meehl, G. A., Covey, C., Taylor, K. E., Delworth, T., Stouffer, R. J., Latif, M., McAvaney, B., and Mitchell, J. F. B. (2007b). The WCRP CMIP3 Multimodel Dataset : A New Era in Climate Change Research. *Bulletin of the American Meteorological Society*, 88(9) :1383–1394.

Murphy, J., Booth, B., Collins, M., Harris, G., Sexton, D., and Webb, M. (2007). A methodology for probabilistic predictions of regional climate change from perturbed

physics ensembles. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 365(1857) :1993–2028.

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., and Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, 430(7001) :768–772. 10.1038/nature02771.

Nakicenovic, N., Davidson, O., Davis, G., Grübler, A., Kram, T., La Rovere, E. L., Metz, B., Morita, T., Pepper, W., Pitcher, H., Sankovski, A., Shukla, P., Swart, R., Watson, R., and Dadi, Z. (2000). *Special Report on Emissions Scenarios : A Special Report of Working Group III of the Intergovernmental Panel on Climate Change.* Cambridge University Press.

Nikiema, O. and Laprise, R. (2011). Diagnostic budget study of the internal variability in ensemble simulations of the Canadian RCM. *Climate Dynamics*, 36(11) :2313–2337.

Peings, Y. and Douville, H. (2010). Influence of the Eurasian snow cover on the Indian summer monsoon variability in observed climatologies and CMIP3 simulations. *Climate Dynamics*, 34(5) :643–660–660.

Peixoto, J. P. and Oort, A. H. (1992). *Physics of Climate.* American Institute of Physics.

Pennell, C. and Reichler, T. (2011). On the effective number of climate models. *Journal of Climate*, (24) :2358–2367.

Pirtle, Z., Meyer, R., and Hamilton, A. (2010). What does it mean when climate models agree? A case for assessing independence among general circulation models. *Environmental Science & Policy*, 13(5) :351–361.

Räisänen, J. (2002). CO2-induced changes in interannual temperature and precipitation variability in 19 CMIP2 experiments. *Journal of Climate*, 15(17) :2395–2411.

Räisänen, J. (2007). How reliable are climate models? *Tellus A*, 59(1) :2–29.

Räisänen, J., Ruokolainen, L., and Ylhäisi, J. (2010). Weighting of model results for improving best estimates of climate change. *Climate Dynamics*, 35(2) :407–422.

Randall, D., Wood, R., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., Stouffer, R., Sumi, A., and Taylor, K. (2007). Climate Models and Their Evaluation. In : *Climate Change 2007 : The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Reif, F. (1965). *Fundamentals of Statistical and Thermal Physics (McGraw-Hill Series in Fundamentals of Physics)*. McGraw-Hill Science/Engineering/Math.

Reifen, C. and Toumi, R. (2009). Climate projections : Past performance no guarantee of future skill ? *Geophys. Res. Lett.*, 36(13) :L13704.

Rowell, D. (2006). A Demonstration of the Uncertainty in Projections of UK Climate Change Resulting from Regional Model Formulation. *Climatic Change*, 79(3) :243–257.

Rowlands, D. J., Frame, D. J., Ackerley, D., Aina, T., Booth, B. B. B., Christensen, C., Collins, M., Faull, N., Forest, C. E., Grandey, B. S., Gryspeerdt, E., Highwood, E. J., Ingram, W. J., Knight, S., Lopez, A., Massey, N., McNamara, F., Meinshausen, N., Piani, C., Rosier, S. M., Sanderson, B. M., Smith, L. A., Stone, D. A., Thurston, M., Yamazaki, K., Hiro Yamazaki, Y., and Allen, M. R. (2012). Broad range of 2050 warming from an observationally constrained large climate model ensemble. *Nature Geosci*, 5(4) :256–260. 10.1038/ngeo1430.

Rummukainen, M. (2010). State-of-the-art with regional climate models. *Wiley Interdisciplinary Reviews : Climate Change*, 1(1) :82–96.

Russell, G. L., Miller, J. R., and Rind, D. (1995). A coupled atmosphere-ocean model for transient climate change studies. *Atmosphere-Ocean*, 33(4) :683–730.

Russell, G. L., Miller, J. R., Rind, D., Ruedy, R. A., Schmidt, G. A., and Sheth, S. (2000). Comparison of model and observed regional temperature changes during the past 40 years. *Journal of Geophysical Research : Atmospheres*, 105(D11) :14891–14898.

Santer, B. D., Mears, C., Doutriaux, C., Caldwell, P., Gleckler, P. J., Wigley, T. M. L., Solomon, S., Gillett, N. P., Ivanova, D., Karl, T. R., Lanzante, J. R., Meehl, G. A., Stott, P. A., Taylor, K. E., Thorne, P. W., Wehner, M. F., and Wentz, F. J. (2011). Separating signal and noise in atmospheric temperature changes : The importance of timescale. *J. Geophys. Res.*, 116(D22) :D22105.

Sato, M., Hansen, J. E., McCormick, M. P., and Pollack, J. B. (1993). Stratospheric aerosol optical depths. *J. Geophys. Res.*, 98(D12) :22987–22994.

Scheffé, H. (1970). Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association*, 65(332) :1501–1508.

Seager, R., Ting, M., Held, I., Kushnir, Y., Lu, J., Vecchi, G., Huang, H.-P., Harnik, N., Leetmaa, A., Lau, N.-C., Li, C., Velez, J., and Naik, N. (2007). Model projections of an imminent transition to a more arid climate in southwestern North America. *Science*, 316(5828) :1181–1184. 10.1126/science.1139601.

Separovic, L., Elía, R., and Laprise, R. (2012). Impact of spectral nudging and domain size in studies of RCM response to parameter modification. *Climate Dynamics*, 38(7-8) :1325–1343.

Sorteberg, A. and Kvamstø, N. G. (2006). The effect of internal variability on anthropogenic climate projections. *Tellus A*, 58(5) :565–574.

Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D. J., Kettleborough, J. A., Knight, S., Martin, A., Murphy, J. M., Piani, C., Sexton, D., Smith, L. A., Spicer, R. A., Thorpe, A. J., and Allen, M. R. (2005). Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, 433(7024) :403–406. 10.1038/nature03301.

Stouffer, R. J. (2004). Time scales of climate response. *Journal of Climate*, 17(1) :209–217.

Stouffer, R. J., Weaver, A. J., and Eby, M. (2004). A method for obtaining pre-twentieth century initial conditions for use in climate change studies. *Climate Dynamics*, 23(3) :327–339.

Tebaldi, C. and Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 365(1857) :2053–2075.

Tebaldi, C., Mearns, L. O., Nychka, D., and Smith, R. (2005a). Regional probabilities of precipitation changes : A bayesian analysis of multi-model simulations. *Geophys. Res. Lett.*, 31. L24213.

Tebaldi, C., Smith, R. L., Nychka, D., and Mearns, L. O. (2005b). Quantifying uncertainty in projections of regional climate change : A bayesian approach to the analysis of multimodel ensembles. *Journal of Climate*, 18(10) :1524–1540.

Tompkins, A. M. (2002). A prognostic parameterization for the subgrid-scale variability of water vapor and clouds in large-scale models and its use to diagnose cloud cover. *Journal of the Atmospheric Sciences*, 59(12) :1917–1942.

van der Linden, P. and Mitchell, J. (2009). Report. ENSEMBLES : Climate Change and its Impacts : Summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK. 160pp (2009).

von Storch, H. and Zwiers, F. W. (1999). *Statistical analysis in climate research.* Cambridge University Press, UK.

Weigel, A. P., Liniger, M. A., and Appenzeller, C. (2008). Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, 134(630) :241–260.

Whetton, P., Macadam, I., Bathols, J., and O'Grady, J. (2007). Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models. *Geophys. Res. Lett.*, 34.

Wigley, T. M. L., Ammann, C. M., Santer, B. D., and Raper, S. C. B. (2005). Effect of climate sensitivity on the response to volcanic forcing. *J. Geophys. Res.*, 110(D9) :D09107.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences, Volume 100, Third Edition.* International Geophysics Series. Elsevier Science & Technology.