A Markov chain method for weighting climate model ensembles and uncertainty estimation on spatially explicit data

Max Kulinich

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Mathematics and Statistics

Faculty of Science

The University of New South Wales

June 2022

Abstract

Climate change is typically modelled using sophisticated mathematical models (climate models) of physical processes that range in temporal and spatial scales. Multi-model ensemble means of climate models show better correlation with the observations than any of the models separately. Currently, an open research question is how climate models can be combined to create an ensemble mean in an optimal way. We present a novel stochastic approach based on Markov chains to estimate model weights in order to obtain ensemble means and uncertainty estimations on spatially explicit climate data. The method was compared to existing alternatives by measuring its performance in cross-validation and model-as-truth experiments on a diverse set of public climate datasets. The Markov chain method showed improved performance over those methods when measured by a set of metrics: root mean squared error, climatological monthly root mean squared error, monthly trend bias, interannual variability, uncertainty error etc. The results of this comparative analysis should serve to motivate further studies in applications of Markov chain and other nonlinear methods that address the issues of finding optimal model weight for constructing weighted ensemble means and uncertainty estimations.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisors Professors Yanan Fan, Spiridon Penev and Jason P. Evans for their invaluable advice, constant support, and incredible patience during my PhD research. Their immense knowledge and dedication encouraged me throughout my academic journey. I would also like to thank Professor Ian Doust, Ms Markie Lugton and Ms Belinda Lee for their help and guidance that made my study and life in Australia an enjoyable and memorable experience. I would also like to acknowledge with gratitude the generous financial support from the UNSW Scientia PhD Scholarship Scheme as this endeavour would not have been possible without it. I am also thankful to Dr. Roman Olson for his contributions to Chapter 2.

I am endlessly grateful to my wife Alena for her unconditional daily support and loving care. My appreciation also goes to my mother and my sisters whose encouragement during my studies helped to keep my spirits and motivation high. Last but not least, I would like to thank my teachers, colleagues and friends in Sydney, Gothenburg and Minsk who influenced and inspired me in many different ways.

Publications and code availability

Publications

- Kulinich, M., Fan, Y., Penev, S., Evans, J. P., and Olson, R.: A Markov chain method for weighting climate model ensembles, Geosci. Model Dev., 14, 3539–3551, https://doi.org/10.5194/gmd-14-3539-2021, 2021.

Code availability

- Code for Chapter 2 is available at https://doi.org/10.5281/zenodo.4548417
- Code for Chapters 3 & 4 is available at https://doi.org/10.5281/zenodo.6698970

Contents

Abbreviations

Abbreviations

AR	Assessment report
AVE	Multi-model ensemble average
CMIP	Coupled Model Intercomparison Project
COE	Convex optimisation ensemble
GCM	Global Climate Model
GPCC	Global Precipitation Climatology Centre
GSAT	Global surface air temperature
HWA	Heatwave amplitude
IPCC	Intergovernmental Panel on Climate Change
KMA	Korea Meteorological Administration
LCL	Lower control limit
MCE	Markov Chain Ensemble
NARCliM	New South Wales (NSW) and Australian Capital Territory Regional Climate Modelling
PI	Prediction interval
RCM	Regional Climate Model
RMSE	Root mean squared error
SSP	Shared Socioeconomic Pathways
UA	Uncertainty area
UCL	Upper control limit

UE	Uncertainty error
VCOE	Varying convex optimisation ensemble
VMCE	Varying Markov Chain Ensemble
WQ	Weighted quantile

Chapter 1

Introduction

1.1 Background and literature review

Earth's climate is a complex system regulated by large-scale interactions between its various components including solar radiation, winds, ocean currents, etc. An important research question in recent years is the influence of anthropogenic greenhouse gas emissions on such climatological characteristics as temperature and precipitation. A common approach to studying this influence has been by developing sophisticated mathematical models of physical processes taking place over a range of temporal and spatial scales. Those climate models provide a crucial source of information about the future climate state allowing planning, mitigating and adapting to the forthcoming climate changes. The most comprehensive assessment of the future climate state is conducted by the Intergovernmental Panel on Climate Change (IPCC) and its 2021 IPCC sixth assessment report (AR6) provides the latest update on future climate projections based on the new generation of climate models from Coupled Model Intercomparison Project (CMIP6)(Lee et al. (2021)).

Even the most sophisticated climate models used for climate change projections are inherently limited in their ability to represent all aspects of the modelled physical processes. They cannot capture all possible system states since any given model is a simplified rep-

resentation of actual climate processes which provides an incomplete though useful description of the world. A common practice to mitigate such climate models' limitations has been to combine several models into a multi-model ensemble. Such ensembles are typically characterised by their best estimations and uncertainty range (or probabilistic distributions). Traditionally, future climate projects are represented by such multi-model ensemble averages (weighted and unweighted) in the research literature and IPCC reports with the spread of the models used as a measure of uncertainty (Stocker et al. (2014)). A simple multi-model mean has been by far the most common and accepted approach to combining models for future climate projections (Lambert and Boer (2001), Gleckler et al. (2008), etc.). Such multi-model means often show better correlations with the observations than any of the individual models separately (Kharin and Zweirs (2002); Feng et al. (2011)). The ensemble multi-model means generally have significantly less variance than the ensemble model outputs individually or the observations (See Figure 1.1 taken from IPCC AR6). Knutti et al. (2010) point out that the multi-model means approach (or equally-weighted average) assumes that all models are (a) reasonably independent, (b) equally plausible, (c) distributed around reality and (d) that the range of their projections is representative of what we believe is the uncertainty in the projected quantity. However, these assumptions are rarely fulfilled (Knutti et al. (2017)), and thus better ways of finding weighted ensemble means were proposed by in more recent studies (e.g. Bishop and Abramowitz (2013), Herger et al. (2018), Sanderson et al. (2017), etc.).

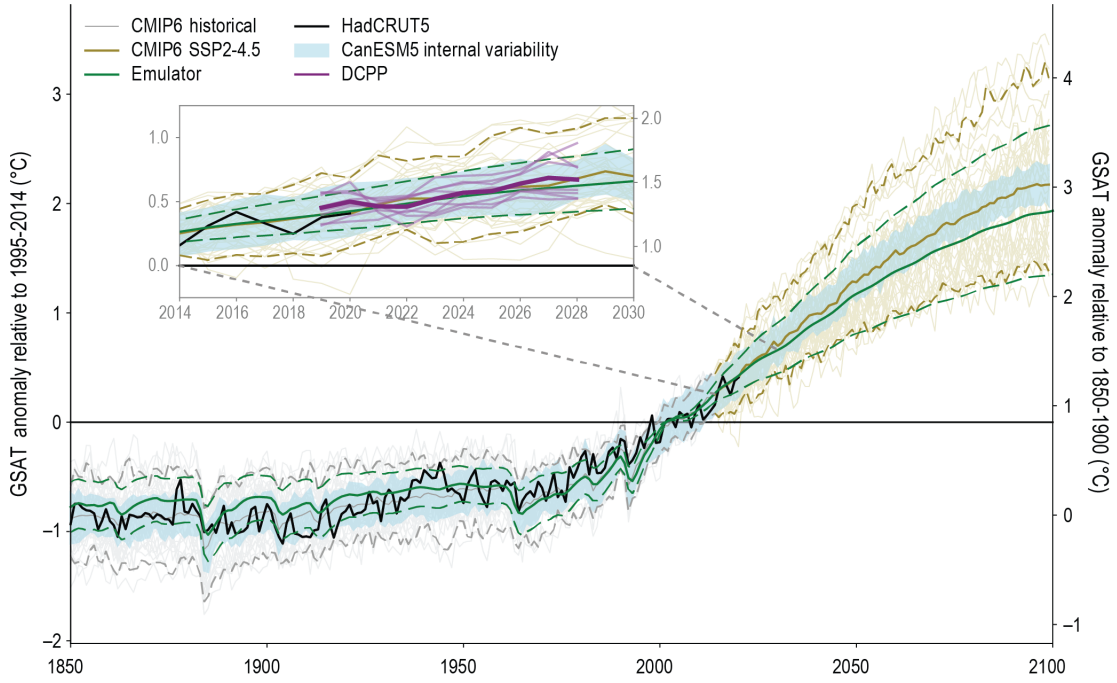

Figure 1.1: CMIP6 annual mean global surface air temperature (GSAT) simulations and various contributions to uncertainty in the projections ensemble. The figure shows anomalies relative to the period 1995–2014 (left y-axis), converted to anomalies relative to 1850–1900 (right y-axis); the difference between the y-axes is 0.85°C. Shown are historical simulations with 39 CMIP6 models (grey) and projections following scenario SSP2-4.5 (dark yellow; thin lines: individual simulations; heavy line; ensemble mean; dashed lines: 5% and 95% ranges). The black curve shows the observations-based estimate (HadCRUT5; Morice et al. (2021)). Light blue shading shows the 50-member ensemble CanESM5, such that the deviations from the CanESM5 ensemble mean have been added to the CMIP6 multi-model mean. The green curves are from the emulator and show the central estimate (solid) and very likely range (dashed) for GSAT. The inset shows a cut-out from the main plot and additionally in light purple for the period 2019–2028 the initialized forecasts from eight models contributing to DCPP (Boer et al. (2016)); the deep-purple curve shows the average of the forecasts (adapted from Figure 1 in IPCC AR6, Chapter 4 (Lee et al. (2021)).

Some of these studies suggest weighting models by applying a static weights vector to (Krishnamurti et al. (2000), Bishop and Abramowitz (2013), Abramowitz et al. (2018), Haughton et al. (2015), Liang et al. (2021) etc.) or sub-selecting (Gleckler et al. (2008), Herger et al. (2018), etc.) different models based on their ability to simulate historical observations. Other studies focus on estimating likelihoods of the model and observation

data (Murphy et al. (2004), Fan et al. (2017), etc.). Such approaches can be described as linear (or based on a generalised linear model) optimisation techniques and are consequently limited by the strong assumptions used for their design.

1.2 Motivation and aims of this thesis

The goal of this thesis is to investigate whether the assumptions used for linear optimisation techniques can be weakened by complimenting the existing climate model weighting methods with a more flexible nonlinear optimisation approach. The application of nonlinear techniques such as data mining on climate data is still in its infancy (Crawford et al. (2019)) and we seek to demonstrate that a robust, consistent nonlinear approach can be developed based on the well-known and widely applied mathematical models such as Markov chains. Provided that nonlinear methods show advantages over linear methods in terms of efficiency, precision, simplicity, interpretability and other desirable properties their application would allow better understanding of climate projections and possibly increase confidence in the prediction of future climate states.

We aim to provide a comprehensive evaluation framework for assessing climate model weighting methods, describe a possible realisation of nonlinear method in the form of Markov chain-based algorithms and compare its efficiency to commonly used linear methods by demonstrating its performance on different types of climate data. In order to evaluate the comparative efficiency of a proposed method for building a climate model ensemble projection a robust set of metrics for its performance is required and the most important current issues with linear models need to be identified. A common and well-discussed disadvantage of a simple multi-model ensemble mean is lack of independence between ensemble members (Knutti et al. (2010), Olson et al. (2019), etc.) since research organisations share climate model software code, model components, model parameters, literature, etc. Abramowitz et al. (2018) point out that model dependence can play a crucial role when assembling the models into an ensemble. Having a large cluster of dependent models in an ensemble with equal weights may result in the overall ensemble

mean being close to this cluster regardless of its modelling skill. Thus, ignoring the model dependence issue can lead to bias and overconfidence in future climate model projections (Leduc et al. (2015), Steinschneider et al. (2015)).

Another important requirement for an ensemble weighting method is its ability to represent not only the best estimation of the future climate state, but provide an uncertainty estimation of many possible future states. As the future climate model projections are affected by a range of uncertainties such as emission scenario uncertainty, climate system's internal variability, model response uncertainty and other factors (Hawkings et al. (2009)) estimating the range of possible outcomes is necessary for interpreting best estimate results. The raw spread of the ensemble models does not necessarily represent the actual climate system's uncertainty (Brunner et al. (2020), Lorenz et al. (2018)) as it is highly dependent on the selection of the models, their interdependence, etc.

As prediction of regional climate change is becoming increasingly important for decision-making and developing mitigation policies (e.g. Christensen et al. (2013), Xie et al. (2015), Almazroui et al. (2021)), an ability to distinguish regional patterns and analyse different climate zones is also required from an ensemble weighting technique. Spatially explicit data can be highly variable across the geographical locations (see Figure 1.2 taken from IPCC AR6) and it presents an additional challenge for climate models and ensemble weighting methods alike.

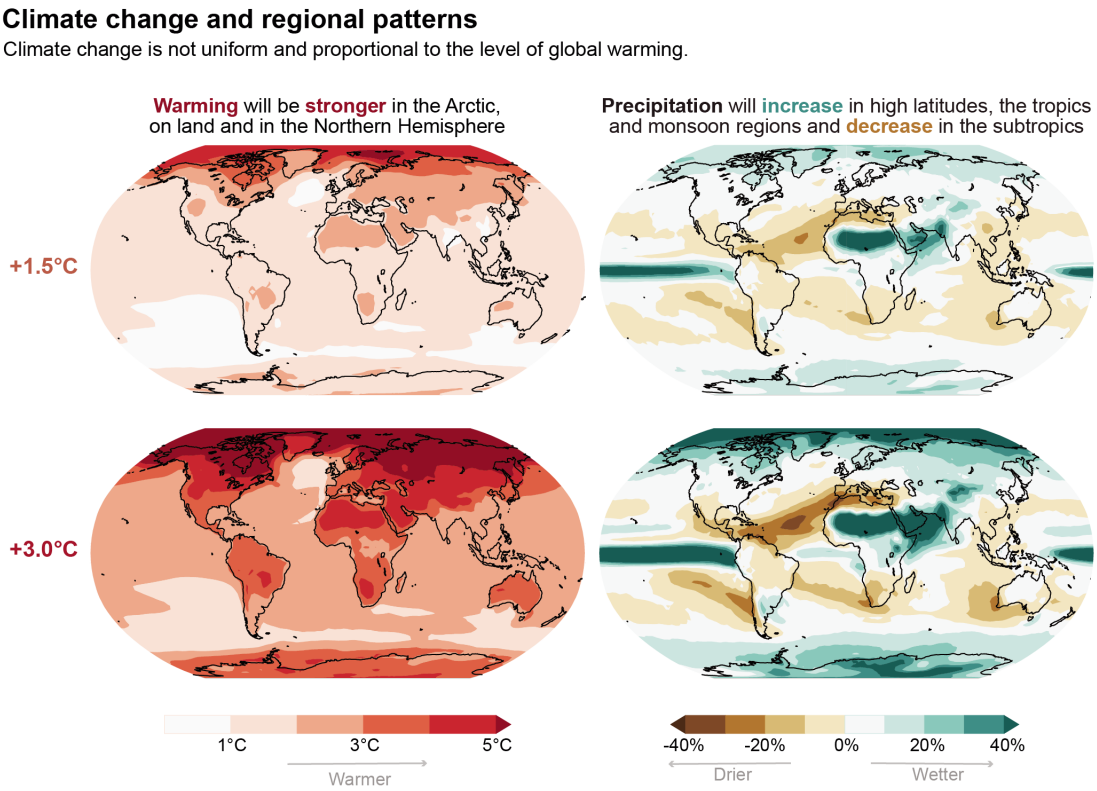

Figure 1.2: Figure 1 | Regional changes in temperature (left) and precipitation (right) are proportional to the level of global warming, irrespective of the scenario through which the level of global warming is reached. Surface warming and precipitation change are shown relative to the 1850–1900 climate, and for time periods over which the globally averaged surface warming is 1.5°C (top) and 3°C (bottom), respectively. Changes presented here are based on 31 CMIP6 models using the high-emissions scenario SSP3-7.0 (adapted from Figure 1 in IPCC AR6, Chapter 4 (Lee et al. (2021)).

Hence, the desirable characteristics of an ensemble weighting method include robustness against the models' interdependence issue, ability to estimate uncertainty, interpretability including using normalised non-negative weights, application flexibility (including spatially explicit data) and high performance across a range of comprehensive statistical and climatology metrics.

To address those desirable characteristics we propose a novel way to construct a weighted ensemble mean using Markov chains, which we call Markov Chain Ensemble (MCE) method. Markov chain method is well-studied and established approach in nonlinear

applications and has some of the desirable characteristics described above embedded in its design.

We aim to demonstrate that going beyond linear optimisation allows better performing weighted ensembles by enhancing the MCE method with a varying weights scheme, which we call Varying Markov Chain Ensemble (VMCE) method. We also examine likelihood interval estimation methods to compliment the best estimates provided by MCE and VMCE with a possible range of future climate states.

1.3 Methods and datasets

Although Markov chains have been used frequently in the literature for the prediction of future time series (e.g. Bai and Wang (2011), Pesch et al. (2015)), to the best of our knowledge, this is the first time this method is being applied to building weighted climate model ensemble means. Markov chain method naturally produces non-negative weights that sum to one and captures some of the nonlinear patterns in the ensemble. We use the the "memoryless" property of Markov chains at each time step to capture the dynamic change in models' fit through the time series. This property becomes particulary advantageous when introducing varying weighting scheme as we will demonstrate below.

We will demonstrate that MCE and VMCE have high relative performance across different datasets, statistical and climatological metrics, geographical locations and climate variables by comparing it to the commonly used multi-model ensemble average (AVE) and convex optimisation ensemble (COE) method similar to the one proposed by Bishop and Abramowitz (2013). In addition we modify COE method to allow for a varying weighting scheme, which we call varying convex optimisation ensemble (VCOE) method, to study the effect of this scheme's application beyond the MCE method.

The rationale behind COE method is to use correlation of model errors as a basis for measuring model dependence and transforming the initial climate model ensemble by applying a vector of weights in order to find a linear combination of ensemble model's

outputs with minimal mean square difference (Bishop and Abramowitz (2013)). This vector of weights is constrained to contain only non-negative values and sum to 1. In this study we implement this approach using an R software package for Disciplined Convex Optimisation (Fu et al. (2020)). As convex optimisation in itself remains outside of this study's scope, we use the recommended default settings proposed by authors without any software modifications to compare its results with AVE, MCE and VMCE methods.

To ensure reliability and consistency of the performance evaluation results we apply the weighting ensemble methods on several well-known climate datasets with temperature, precipitation and heatwave amplitudes data. To study the performance dependency on geographical location we apply all the methods on all land points of a 144 by 72 global grid. To cover a comprehensive range of statistical and climatological performance metrics we include root mean square error, climatological monthly root square error, climatological monthly bias, trend bias and interannual variability bias into the analysis. Furthermore we use cross-validation procedure and model-as-truth (out-of-sample) performance assessment to ensure the objective evaluation of possible future climate estimation skill for each method.

We introduce a new way of uncertainty estimation which is different from common methods based on weighted and unweighted quantiles of climate models' outputs. We demonstrate that this novel uncertainty estimation method based on prediction interval commonly used in regression analysis is less prone to underestimating and overestimating uncertainty of climate projections within a calibration/validation framework. Finally, we summarise all the findings to estimate global annual climate change by calculating best estimates and prediction intervals for the introduced MCE method and compare those results with the commonly used AVE method.

1.4 Thesis outline

The structure of this thesis is as follows. In Chapter 2 we present an overview of studies attempting to define an optimal ensemble weighting method, current unresolved issues in using weights, desirable properties for an ensemble weighting method and a novel way to construct a weighted ensemble mean using an approach based on Markov chains. We demonstrate its performance on three distinct climate datasets for a range of statistical and climatological metrics and compare the results to AVE and COE methods. We discuss MCE method's advantages and limitations and how it mitigates the climate models' dependability issue.

In Chapter 3 we introduce a varying weight Markov chain ensemble (VMCE) method and apply it together with AVE, COE, MCE and a varying weight convex optimisation ensemble (VCOE) methods on spatially explicit climate data. We demonstate VMCE performance on two distinct climate variables (temperature and precipitation) for a range of statistical and climatological metrics and analyse geographical patterns of the results. We summarise the findings in a global annual climate projections for temperature and precipitation and discuss advantages and limitations of VMCE method.

In Chapter 4 we introduce a novel uncertainty estimation method based on a prediction interval approach. We apply it for all the weighted ensemble means (AVE, COE, MCE, VCOE and VMCE) on spatially explicit climate data (temperature and precipitation) and compare it to a weighted quantile method by measuring its error and uncertainty area. We summarise the findings in a global annual climate uncertainty projections for temperature and precipitation and discuss advantages and limitations of the prediction interval application to climate data.

In Chapter 5 we summarise all the previous findings into overall thesis conclusions and provide a guidance for future research directions.

Chapter 2

A Markov chain method for weighting climate model ensembles

In Chapter 2, we will present an overview of studies attempting to define an optimal ensemble weighting, current unresolved issues in using weights, desirable properties for an ensemble weighting method and a novel way to construct a weighted ensemble mean using an approach based on Markov chains, which addresses those issues discussed. To our knowledge there were no applications of Markov chain method for constructing climate models' weighted ensemble mean. We describe the mathematical foundation of the method, data used for experiments, performance metrics designed to evaluate the efficiency of different ensemble weighting methods, cross-validation techniques and results obtained, followed by a discussion about advantages and limitations of the Markov chain approach in comparison to other existing methods.

This work has been accepted for publication in:

Kulinich, M. and Fan, Y. and Penev, S. and Evans, J. P. and Olson, R., A Markov chain method for weighting climate model ensembles, Geosci. Model Dev., 14, 3539–3551, 2021, https://doi.org/10.5194/gmd-14-3539-2021

Max Kulinich with contributions from all co-authors developed the methods' theoretical framework and code, designed the experiments, performed the simulations and prepared the paper. He prepared CMIP5 data and Roman Olson prepared NARCLiM and KMA data.

2.1 Introduction

Climate change is often modelled using sophisticated mathematical models of physical processes taking place over a range of temporal and spatial scales. These models are inherently limited in their ability to represent all aspects of the modelled physical processes. Simple averages of multi-model ensembles of GCMs (Global Climate Models) often show better correlations with the observations than any of the individual models separately (Kharin and Zweirs (2002); Feng et al. (2011)). Knutti et al. (2010) point out that often the equal-weighted averages ("one model, one vote") approach is used as a best-guess result, assuming that individual model biases will at least partially cancel each other out. This approach assumes that all models are (a) reasonably independent, (b) equally plausible, (c) distributed around reality and (d) that the range of their projections is representative of what we believe is the uncertainty in the projected quantity. However, these assumptions are rarely fulfilled (Knutti et al. (2017)), and thus a better way of finding a weighted ensemble mean is required (Herger et al. (2018); Sanderson et al. (2017)).

Most studies attempting to define an optimal ensemble weighting either employ linear optimisation techniques (Abramowitz et al. (2018); Krishnamurti et al. (2000); Majumder et al. (2018)) or are based on a specification of likelihoods for the model and observation data (Fan et al. (2017); Murphy et al. (2004)). Such methods are inevitably limited by the strong assumptions used for their design. We seek to weaken those assumptions and to complement the existing methods with a more flexible nonlinear optimisation approach. An unresolved issue in using weights for models is that models have interdependence, due to the sharing of computer codes, parameterizations, etc. (Olson et al. (2019)). Abramowitz et al. (2018) points out that model dependence can play a crucial role when assembling

the models into an ensemble. Mathematically, interdependence often result in closeness of model outputs in model output space. If a large cluster of highly dependent models is included into an ensemble with equal weights, the overall ensemble mean will become close to the dependent models' cluster. Ignoring model dependence can lead to bias and overconfidence in future climate model projections (Leduc et al. (2015); Steinschneider et al. (2015)).

Hence, it is desirable that an ensemble weighting method is robust against the dependency issue, and has normalised non-negative weights for interpretability. Finally, the methods should work well across a range of different climate variables, such as temperature, precipitation, etc.

In this paper, we propose a novel way to construct a weighted ensemble mean using Markov chains, which we call the Markov Chain Ensemble (MCE) method. Our purpose is to demonstrate that going beyond linear optimisation on a vector space of climate models' outputs allows building better performing weighted ensembles. We selected Markov chains as a basis for such nonlinear optimisation as one of the most straightforward nonlinear structures. It naturally produces non-negative weights that sum to one and captures some of the nonlinear patterns in the ensemble (here we refer to nonlinear patterns as time-dependent selection of model components rather than considering complete model output vector). It performs well on a range of datasets when compared to the standard simple mean and linear optimisation weighting methods as we demonstrate below. We also examine how the method responds to the introduction of interdependent models.

Although Markov chains have been used frequently in the literature for the prediction of future time series (e.g. Bai and Wang (2011); Pesch et al. (2015)), to the best of our knowledge, this is the first time the method has been applied to building weighted climate model ensemble means. In this paper, we use the "memoryless" property of Markov chains at each time step to capture the dynamic change in models' fit through the time series. This dynamic change, through time, is represented by the transition matrix, which describes the probability of each model being the best fit for the next observation at time $t+1$, given the best fit for the current time t. The transition matrix is built based on the

input data and describes probable future states given the current state. The stationary distribution of this transition matrix is used for weighted ensemble creation and reflects the relative contribution of each model to the total weighted ensemble mean forecast.

We describe the datasets used in this study and the proposed MCE method in Section 2. We compare the proposed method (MCE) to the commonly used multi-model ensemble average (AVE) method (Lambert and Boer (2001)) and the convex optimisation (COE) method proposed by Bishop and Abramowitz (2013) and present the results in Section 3, followed by a discussion in Section 4 and conclusion in Section 5.

2.2 Methods

2.2.1 Data

Here we first describe the datasets used in this study. We have chosen three publicly available datasets with differing number of models, historical period lengths and model interdependence levels to evaluate and compare the performance of the MCE method with alternative approaches.

CMIP5 Data: The first dataset we use is the temperature anomalies (°C) data from Coupled Model Intercomparison Project (CMIP5) with 39 different Global Climate Model (GCM) outputs (one ensemble member per model) and Hadley Centre/Climatic Research Unit Temperature observations (HadCRUT4). The data is obtained from `https://climexp.knmi.nl` and the period of 1900 - 2099 is selected for the analysis. It contains temperature anomalies (monthly averages) compared to the reference period of 1961-1990 (Taylor et al. (2011)). This dataset contains several clusters of dependent models, has both positive and negative data values, a relatively low variability and long time series.

NARCliM Data: The second dataset contains temperature output from the New South Wales (NSW) and Australian Capital Territory Regional Climate Modelling project (Evans

et al. (2014)). It contains regional climate model (RCM) simulations over southeastern Australia. Specifically, three RCM versions were forced with four global climate models each, for a total of twelve ensemble members. The data contains annual time series of mean summer temperature (°C) for the Far West NEW state planning region as modelled by the NARCliM domain regional climate models (RCMs) for the periods 1990–2019 and 2030-2039 (Olson et al. (2016)). Corresponding temperature observations are obtained from the Australian Water Availability Project (AWAP) (Jones et al. (2009)). The dataset has a high ratio of the number of models to the number of observations. While NARCliM model choice explicitly considered model dependence for both the RCMs as well as the driving GCMs, the resulting ensemble demonstrates an apparent similarity between the simulations (i.e., model inter-dependence) in small clusters.

KMA Data: The third dataset contains yearly heatwave amplitudes (HWA) for the Korean peninsula from 29 CMIP5 climate models and observations between years 1973 and 2005 (Shin et al., 2017). In particular, HWA contains the difference between the highest temperature during the heatwave events for the corresponding year and the 95th percentile of daily maximum summer temperatures from 1973 to 2005. This framework was discussed in detail in Fischer and Schär (2010). Here a heatwave event occurs when the daily maximum temperature is above the 95th percentile of daily maximum summer temperatures (32.82°C) for two consecutive days. Daily maximum temperature data used for the calculation of observed HWA is the mean of 59 weather stations operated by the Korea Meteorological Administration (KMA). Shin et al. (2017) provides the list of CMIP5 models included in the study. HWA data is non-negative and can be highly skewed with long upper tails as it measures extreme events; therefore, the dataset is highly non-Gaussian. These properties allow us to test methods in more challenging scenarios, where likelihood-based approaches are more difficult to apply.

These three datasets cover different scenarios, data structures, parameter distributions and scales (see Table 2.1) . Such coverage allows us to analyse the performance and the inherent limitations of the proposed method. In this pilot study we use spatially averaged data, which limits physical interpretability of the model weights, but the method can be

extended to spatially distributed data.

Dataset	Climate variable	Minimum	Maximum	Variance	Number of observations	Number of models
CMIP5	Temperature (°C)	-0.80	1.16	0.12	1440	39
NARCLiM	Temperature (°C)	9.38	31.64	36.61	240	12
KMA	HWA (°C)	0	1.49	0.18	33	29

Table 2.1: Summary of CMIP5, NARCLiM and KMA data properties.

2.2.2 Markov chain ensemble (MCE) method

Generally, a homogeneous Markov chain is a sequence of random system states evolving through time, where each next state is defined sequentially based on its predecessor and predefined transition probabilities (Del Moral and Penev, 2016, p. 121). Suppose that there is a finite number of probable system states $S = \{s_1, \ldots, s_N\}$, then this dependency can be described through a transition matrix P (with $P(x, y) \in [0, 1]$ and $\sum_y P(x, y) = 1$, for any $x, y \in S$):

$$\forall x, y \in S, \quad Pr(X_{n+1} = y | X_n = x) = P(x, y). \tag{2.1}$$

In this study, we want to utilise the "Fundamental Limit Theorem for Regular Chain" which states that if P is a transition matrix for a regular Markov chain (where $\forall x, y \in S$, $P(x, y) > 0$), then $\lim_{n\to\infty} P^n = P^\infty$ where P^∞ is a matrix with all rows being equal and having strictly positive entries.

This property allows us to construct a non-negative transition matrix P by distillation of input information (i.e., model outputs and historical observations) and allows P to converge to a unique vector of model weights $w = (w_1, w_2, ..., w_N)$, where N is a total number of models in a given ensemble. The vector w can be obtained by solving the equation $wP = w$. The converged transition matrix represents a probability of selecting

one of the models for any of the time steps in the future when observations are not available. Hence, we propose to use it as a weighting vector for constructing a weighted ensemble mean forecast and test this proposition using cross-validation in the following sections.

More precisely, we start by constructing a transition vector v (based on the input data) which specifies a choice of the optimal model at any given time step t. Using the vector v we construct a transition matrix P and find its stationary distribution w. The resulting weighted ensemble mean is constructed by applying w on the given climate model outputs. We call this process Markov Chain Ensemble (MCE) algorithm, and it uses historical observations and equivalent climate model simulations as the input data to calculate a set of weights for the future ensemble mean as an output. Table 2.2 gives a step by step description of the MCE algorithm.

We provide some details of the algorithm as described in Table 2.2 in the following paragraph.

Initialisation of transition matrix P^0: In order for Markov chain to be regular we set $P^0(x, y) = \lambda$, $\forall x, y \in S$, where λ equals the lowest computationally possible positive number $\lambda = 2.225074e^{-308}$ in the R software (R Core Team (2013)).

Initialisation of σ interval: To avoid division by 0 in Equation 2.2 and to prevent Equation 2.2 from converging to $1/N$ the initial σ interval is set to $[0.1, 1]$.

Step 1: The MCE method proceeds by utilising each model output in an optimal way based on its ability to resemble observational data at each given time point. This resemblance is measured by a distance-based probability matrix D of size $N \times T_1$, using a normalised exponential function.

Input:

- length of training period T_1, and
- historical observations O_t, at times $t = 1, \ldots, T_1$, and
- climate model output $M_{i,t}$, at times $t = 1, \ldots, T_1$, for $i = 1, \ldots, N$ models, and
- an initialised number of simulations L
- an initialised σ interval $[\sigma_{min}, \sigma_{max}]$
- an initialised transition matrix P^0 of N x N size

Step 1. Randomly select $\sigma \in [\sigma_{min}, \sigma_{max}]$ and compute the distance matrix D according to Equation 2.2.

Step 2. Construct a sequence vector v based on D using stochastic simulations.

Step 3. Update P^0 step-wise by increasing probability of transitions contained in v : $P^0 \rightarrow P^1 \rightarrow ... \rightarrow P^{T_1}$.

Step 4. Obtain normalised transition matrix P^*, by normalising P^{T_1} row-wise so that each row sums to 1.

Step 5. Find w by solving $wP^* = w$ and store its value.

Step 6. Construct the ensemble mean based on weights w and calculate its $RMSE_{Tr}$

Step 7. Repeat Step 2 - 6 until L sets of weights $w^1, w^2, ..., w^L$ and respective $RMSE^1_{Tr}, RMSE^2_{Tr}, ..., RMSE^L_{Tr}$ have been obtained.

Step 8. Select a set of weights w^* corresponding to the minimal $RMSE^*_{Tr}$

Step 9. Construct the final E_{MCE} using the selected w^*

Table 2.2: The Markov Chain Ensemble (MCE) algorithm.

$$d_{i,k} = \frac{e^{-\left(\frac{M_{i,k}-O_k}{\sigma}\right)^2}}{\sum_{j=1}^{N} e^{-\left(\frac{M_{j,k}-O_k}{\sigma}\right)^2}} \tag{2.2}$$

where $1 \leq k \leq T_1 \leq T$, T_1 indicates the length of the training period, and T is the length of the entire historical period included in the study. Additionally, $1 \leq j \leq N$ where N is the number of models included, and σ is chosen randomly as described above.

Step 2: Based on the matrix D a simulation is performed at each time step $1 \leq k \leq T_1$

by randomly selecting one of the models i with probability proportional to its value $d_{i,k}$. This way we construct a vector $V = (v_1, v_2, v_3, ..., v_{T_1})$, which represents choice of models closest to observations at each time step.

Step 3: Then the initial matrix P^0 is updated step-wise $(P^1, P^2, ..., P^{T_1})$ to capture the transitions between models present in vector V. For each t $(1 \leq t \leq T_1 - 1)$, $P^i_{V_i,V_{i+1}} = P^{i-1}_{V_i,V_{i+1}} + 1$.

Step 4: The resulting matrix is normalised by row $P^*_i = P^{T_1}_i / \sum_{j=1}^{N} P^{T_1}_{i,j}$, for each $1 \leq i \leq N$.

Step 5: The stationary distribution w is obtained by solving $wP^* = w$. A standard R software package is used to find the solution in this study.

Step 6: Construct the ensemble mean based on weights w and calculate its $RMSE_{Tr}$.

Step 7: Steps 2 - 6 are repeated L times, where L is selected based on the external requirements on precision of the results and on computational power available.

Step 8. Select the set of weights w^* with the best performance on the training set (with the lowest $RMSE^*_{Tr}$).

Step 9. Construct the E_{MCE} ensemble using the selected w^*.

2.2.2.1 Parameter sensitivity

From Equation 2.2 it is clear that having a small σ will result in distances d close to $1/N$. Having a large σ will result in all the distances becoming marginal with the exception of the largest one. To optimize the properties of the simulations we control σ by randomly choosing it from $[0.1, 1]$ interval.

As we select only one of the simulations, the MCE method is not sensitive to the number of simulations L after a certain threshold. This threshold is set based on the requirements for precision of the results and on the calculation time. In Figure 2.1 we illustrate the

simulation performance dynamics (simulation index and performance on training and validation NARCLiM data) depending on value of $L \in [1, 1000000]$. The simulation index $i^* \leq L$ represents the index of the best performing simulation at each value of L (with w^* vector of weights and $RMSE^*_{Tr}$ as descibed in Step 8 of Table 2.2). The cross-validation procedure and RMSE metrics are described below in Sections 2.5 - 2.6.

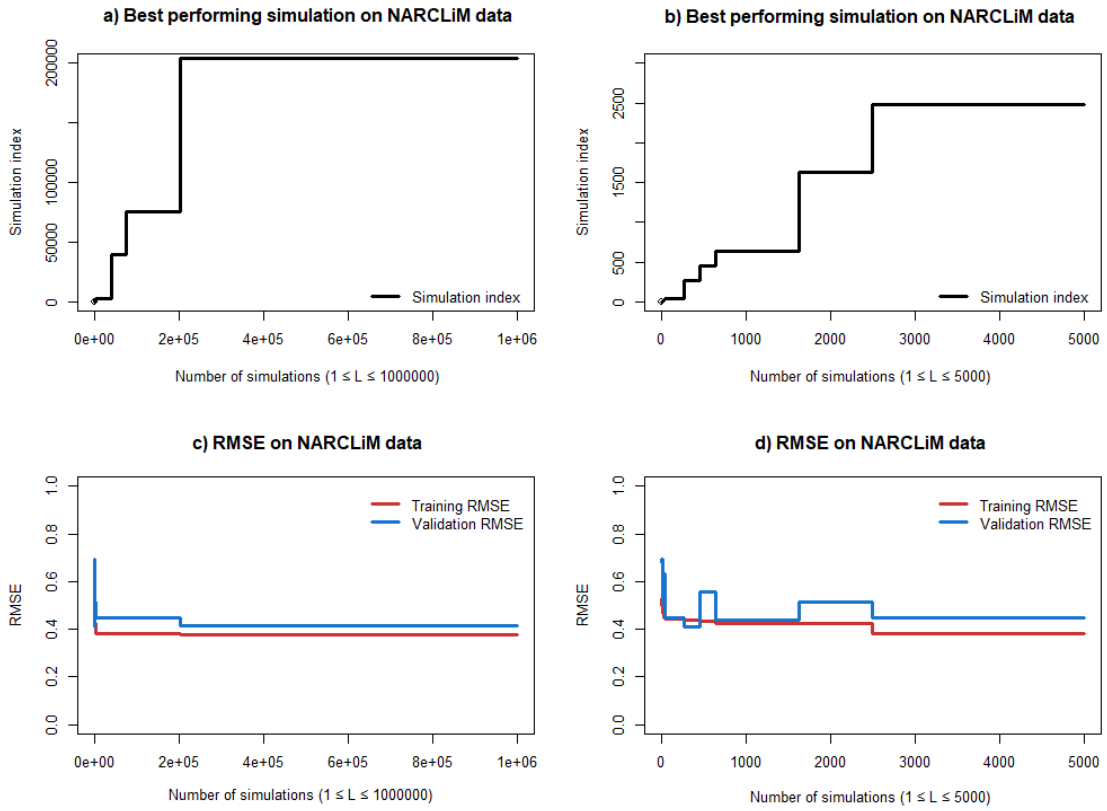

Figure 2.1: Sensitivity of the ensemble properties to the value of L. Left panes a) and c) contain results from all the simulations. Right panes b) and d) contain the results from the first 5000 simulations.

Though better RMSE results can be achieved with larger L, the marginal improvement in RMSE has high computational time cost. For the demonstration purposes in this study we select $L = 3000$ to accommodate for possible differences in RMSE changes between different datasets. As we will show below even with a sub-optimal value of L, MCE method has high performance and stable results.

2.2.2.2 Model interdependence

While we do not claim that the proposed method explicitly addresses the issue of model dependence, it is implicitly addressed to some degree at Step 3 in Table 2.2 of the MCE method. If there are two or more highly correlated models only one of them can be chosen at each step, and thus the resulting sum of such models' weights will be close to the scenario when only one of those models is kept in the ensemble.

We demonstrate this property of the MCE method on modified NARCliM data by adding a copy of one of the models with an added small random error and comparing the resulting weights as shown in Figure 2.2. To mitigate difference in weight values between random simulations we repeat the calculation 100 times and compare the mean values of the weights.

As we can see from Figure 2, adding a highly correlated ensemble member does not significantly change the weights distribution significantly, and more pleasingly when a high performing model is duplicated, the weights are shared between the two copies (see Model 3 and Model 9). Consequently the performance of E_{MCE} remains approximately the same. Though we can not guarantee this behaviour in all types of data, we believe that the MCE method's design helps to mitigate the model interdependence problem.

2.2.2.3 MCE method limitations

Though the MCE method can be used on any climate dataset which contain the required inputs, its relative performance differs depending on the properties of the dataset. We will demonstrate that in the case of a normally distributed data, its performance is competitive with the simple averaging and other more sophisticated methods. In more challenging scenarios, when data is not normally distributed, MCE is performing better than the common alternatives.

As the MCE method is based on a stochastic process, the results between runs can vary.

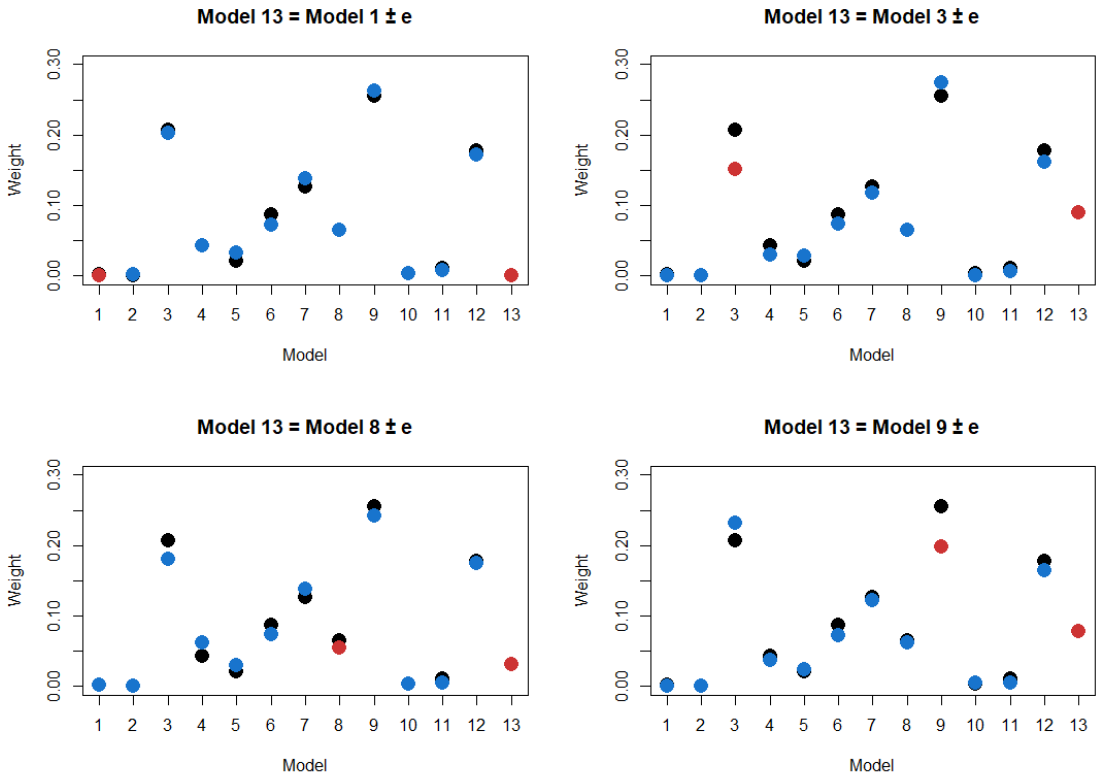

Figure 2.2: Change of MCE weights after adding a copy of Model 1, Model 3, 8 and 9 (clockwise from top left) to the NARCliM ensemble. The original MCE weights are in black. The weights of the modified ensemble are in blue, and the weights of the highly correlated models are in red.

To mitigate this effect and to have reproducible results we set the seed of R software's random number generator to a constant for all simulations. The MCE method in its current implementation does not provide an uncertainty quantification, and this limitation is a subject for future nonlinear ensemble weighting methods development.

Finally, as the MCE method does not consider spatial information, the resulting weights have limited physical interpretability. Extending the MCE method to utilize such information is a subject for future research.

2.2.3 Multi-model ensemble average (AVE) method

In order to evaluate the relative performance of the MCE method we select two other popular approaches to constructing ensemble weighted average. The first approach is the widely used average of individual climate model outputs (Gleckler et al. (2008); Lambert and Boer (2001)):

$$E_{AVE_t} = 1/N \sum_{j=1}^{N} M_{j,t}, \tag{2.3}$$

for each $1 \leq t \leq T$. If model differences from observations are random and independent, they will cancel on averaging and the resulting ensemble average will perform better than individual climate models (Lambert and Boer (2001)).

2.2.4 Convex optimisation (COE) method

The second approach that has been selected for relative performance evaluation in this study is a convex optimisation as proposed by Bishop and Abramowitz (2013). It represents a family of other methods based on a linear optimisation over the vector space of individual climate model outputs.

The purpose of this method is to find a linear combination of climate model outputs with $w_1, w_2, ..., w_N$ weights which would minimise mean squared differences with respect to observations:

$$E_{COE_t} = \sum_{j=1}^{N} w_j M_{j,t}, \tag{2.4}$$

for each $1 \leq t \leq T$, so that $\sum_{t=1}^{T}(E_{COE_t} - O_t)^2$ is minimised under restrictions $\sum_{j=1}^{N} w_j = 1$ and $w_j \geq 0$ for each $1 \leq j \leq N$.

This method and its implementation are discussed in details in Bishop and Abramowitz (2013), and we show that it has relatively high performance on the chosen datasets. However, like any other linear optimisation technique, it naturally has some limitations that nonlinear optimisations like the MCE method do not. In particular, the COE method assumes having a large enough sample size to rule out spurious fluctuations in the weights

associated with too small sample size. Such an assumption is not required for the MCE method. In addition, convex optimisation tends to set a large portion of weights equal to 0, as is shown in the examples below, which results in lower effective number of models used for prediction.

2.2.5 Performance metrics

2.2.5.1 RMSE

The root mean squared error (RMSE), Equation 2.5 is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. RMSE is positive, and a value of 0 indicates a perfect fit to the data. In general, a lower RMSE is better than a higher one. However, comparisons across different types of data would be invalid because the measure is dependent on the scale of the numbers used. Minimising RMSE is commonly used for finding optimal ensemble weight vectors (e.g. Herger et al. (2018); Krishnamurti et al. (2000)).

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{j=1}^{N} w_j M_{j,t} - O_t\right)^2}, \tag{2.5}$$

with $\sum_{j=1}^{N} w_j = 1$ and $w_j > 0$ for $j = 1, \ldots, N$. T is the total number of time steps, $M_{j,t}$ denotes the value of model j at time step t and O_t is the observed value at point t.

2.2.5.2 Trend bias

The monthly trend bias is calculated as the difference between the inclination parameter a in weighted ensembles and observations estimated using a linear function $y = ax + b$ on validation data for each month. The total weighted ensemble trend bias metric is calculated as a mean of the monthly trend biases.

2.2.5.3 Climatology monthly bias

The monthly bias is calculated as the difference between the mean of the weighted ensemble and the observation for each month on validation data. The total climatology monthly bias metric is calculated as a mean of the monthly biases.

2.2.5.4 Interannual variability

Interannual variability for each month is calculated as the difference between the standard deviation of detrended weighted ensemble and the standard deviations of detrended observations on validation data. The total interannual variability metric is calculated as the mean of interannual variability for each month.

2.2.5.5 Climatological monthly RMSE

Climatological RMSE is calculated according to Equation 2.5 on climatological monthly means of weighted ensemble values and observations on validation data.

2.2.6 Cross-validation procedures

2.2.6.1 Holdout method

In this method the dataset, which contains the observations, is split into a training (or calibration) set and a validation (or testing) set. The goal of cross-validation is to examine the model's ability to predict new data that was not used in estimating the required parameters.

We partition our data into two sets, with 70% of data used for training and 30% for validation. This is a specific case of the K-fold validation procedure (Refaeilzadeh et al.,

2009, p. 532-538), which is relatively simple to apply and discuss, facilitating the sharing of our findings with other members of the research and non-research communities.

2.2.6.2 Model-as-truth performance assessment

To evaluate each method's performance on the future model projections, we use the model-as-truth approach and analyse the metrics described in Section 2.5. At each step of model-as-truth performance assessment one model is selected as a true model (pseudo-observations) and the remaining models are used to build a weighted ensemble mean that best estimates the true model over the historical period. This weighted ensemble mean is then tested against the future projections of the true model. For a given ensemble this is repeated as many times as the number of the ensemble members with a different member being chosen as the true model each time. The median and spread of these results is reported.

2.3 Results

2.3.1 CMIP5 data

Though the selected monthly CMIP5 data contains annual variation, it is not predominant due to the length and trend of the dataset as shown in panel a) in Figure 2.3. The CMIP5 models output distribution is close to normal as shown in panel b) in Figure 2.3.

Applying the MCE method on the selected CMIP5 data with $T = 120$ (1900 - 2019) and a training period $T_1 = 80$ (1900 - 1979), we obtain a weighted ensemble mean E_{MCE} and compare it with outputs from other methods. We summarize CMIP5 data properties together with the resulting ensemble's weights in Figure 2.3 and holdout cross-validation results in Table 2.3.

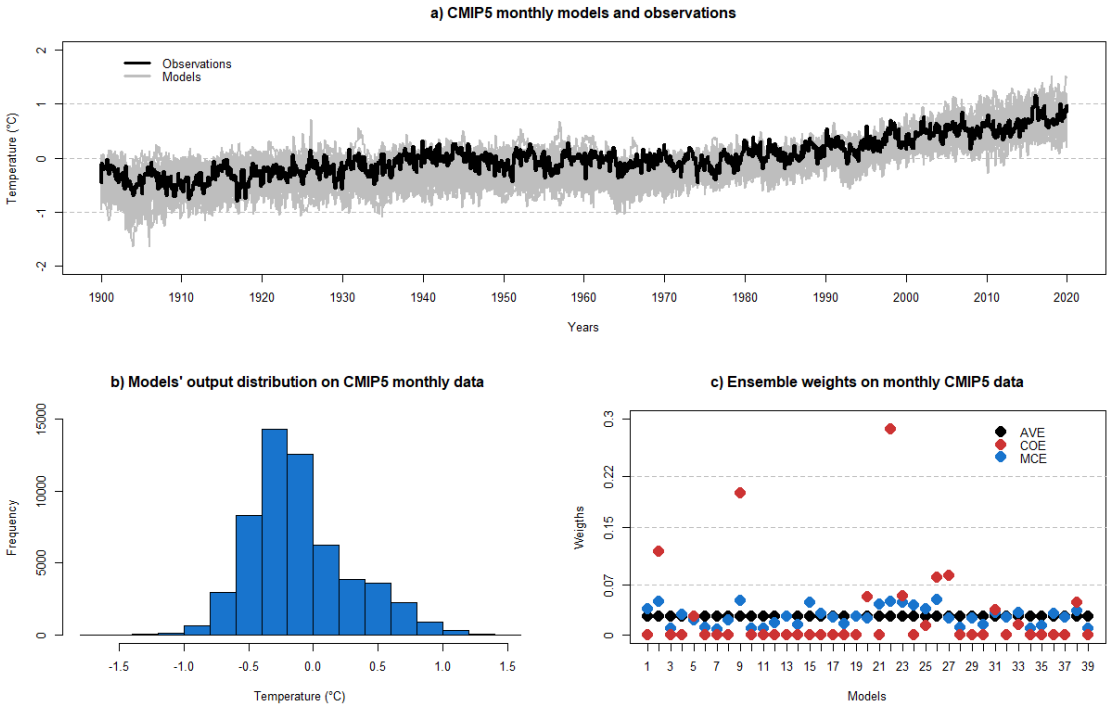

Figure 2.3: CMIP5 data properties. **a)** Model outputs and observations. **b)** Model output distribution. **c)** AVE, COE and MCE weights.

Ensemble	$RMSE_T$	$RMSE_V$	B_T	B_{CM}	B_{IV}	$RMSE_{CM}$
E_{AVE}	0.22	0.17	0.01	-0.09	-0.05	0.10
E_{COE}	0.15	0.19	0.00	-0.12	-0.04	0.13
E_{MCE}	0.18	0.17	0.01	-0.10	-0.05	0.10

Table 2.3: Performance comparison of different methods on CMIP5 data, RMSE on training ($RMSE_T$) and validation ($RMSE_V$) data; trend bias (B_T), climatological monthly bias (B_{CM}), interannual variability bias (B_{IV}) and climatological monthly RMSE ($RMSE_{CM}$) on validation data.

We can see that E_{AVE} and E_{MCE} perform at a similar RMSE level, with E_{COE} performance decreasing comparatively more in validation, a possible indication of overfitting to the training data. We can see from Figure 2.3 that the COE method tends to set zero weights to some models, but builds a weighted ensemble mean that performs best on the training period (1900-1979). Due to some models having zero weights, some of the models' diversity is lost, and this results in worse performance on the validation period ($RMSE_V$ and $RMSE_{CM}$ in Table 2.3). The MCE method, on the other hand, produces model

weights that vary around $1/N$, where N is the number of the models. The MCE method does not give any model zero weighting and hence preserves the ensembles' diversity. The climatological biases B_T, B_{CM} and B_{IV} are nearly equal for all three methods.

The model-as-truth performance assessment is done on $T = 200(1900 - 2099)$ and a training period $T_1 = 120(1900 - 2019)$ as described in Section 2.6.2. The results are summarized in Figure 2.4 and Table 2.4 in form of median, 25% and 75% percentiles of the $N = 39$ (number of models) values.

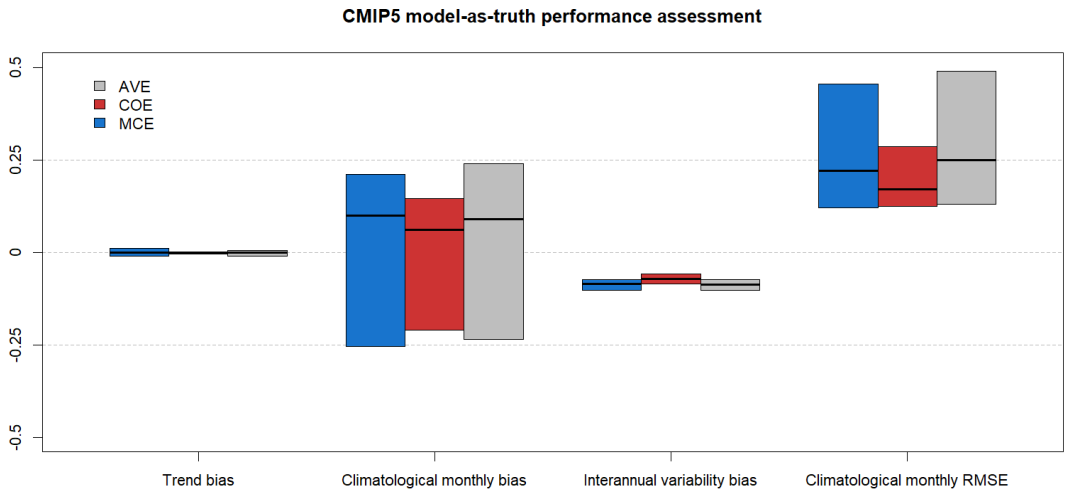

Figure 2.4: CMIP5 model-as-truth performance assessment results. Median, 25% and 75% percentiles of $N = 39$ models.

Ensemble	B_T	B_{CM}	B_{IV}	$RMSE_{CM}$
E_{AVE}	0.00	0.09	-0.09	0.25
E_{COE}	0.00	0.06	-0.07	0.17
E_{MCE}	0.00	0.10	-0.09	0.22

Table 2.4: Model-as-truth performance comparison of different methods on CMIP5 data, median of trend bias (B_T), climatological monthly bias (B_{CM}), interannual variability bias (B_{IV}) and climatological monthly RMSE ($RMSE_{CM}$) on validation data.

All the methods perform similarly in model-as-truth assessment with E_{COE} having better $RMSE_{CM}$.

2.3.2 NARCliM data

The seasonal variation in NARCLiM data is larger than in CMIP5 data as shown in panel a) in Figure 2.5. The NARCLiM models output distribution is not normal as shown in panel b) in Figure 2.5 due to summer time and winter time temperature peaks.

We apply the MCE method on the selected NARCliM data with $T = 20$ (1990 - 2009) and a training period $T_1 = 14$ (1990 - 2003), obtain a weighted ensemble mean E_{MCE} and compare it with outputs from other methods. We summarize NARCLiM data properties together with the resulting ensemble's weights in Figure 2.5 and holdout cross-validation results in Table 2.5.

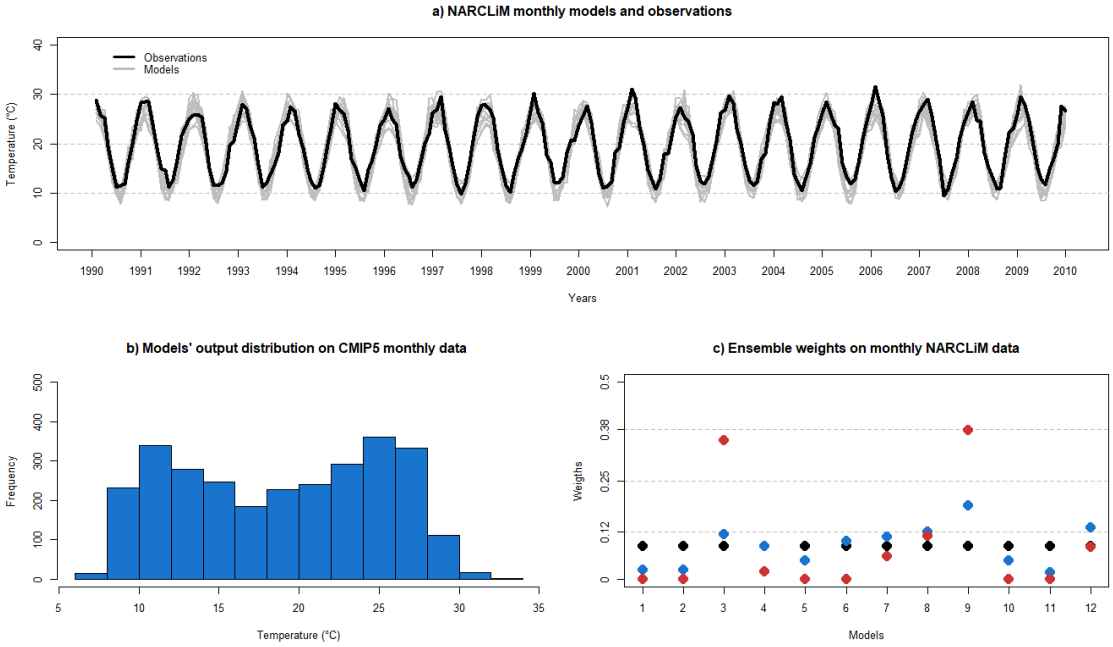

Figure 2.5: NARCLiM data properties. **a)** Model outputs and observations. **b)** Model output distribution. **c)** AVE, COE and MCE weights.

As in CMIP5 data analysis (Figure 2.3), we see that the MCE method is maintaining (i.e., assigning non-zero weights to) more models in the final weighted ensemble than the COE method. As the number of models is significantly smaller than in CMIP5 case, the difference between the MCE output weights and the equal weights is also considerably

Ensemble	$RMSE_T$	$RMSE_V$	B_T	B_{CM}	B_{IV}	$RMSE_{CM}$
E_{AVE}	1.6	1.85	0.04	-1.16	-0.73	1.19
E_{COE}	1.32	1.58	0.00	-0.49	-0.59	0.64
E_{MCE}	1.4	1.58	0.04	-0.64	-0.69	0.70

Table 2.5: Performance comparison of different methods on NARCLiM data, RMSE on training ($RMSE_T$) and validation ($RMSE_V$) data; trend bias (B_T), climatological monthly bias (B_{CM}), interannual variability bias (B_{IV}) and climatological monthly RMSE ($RMSE_{CM}$) on validation data.

larger. The MCE method shows itself capable of maintaining much of the ensembles' diversity during the optimization process. This allows MCE to substantially improve performance over the AVE method on both training and validation periods and perform at the same level as COE on validation period even with lower $RMSE_T$. Again, COE has a larger decline in performance from training to validation periods indicating possible overfitting.

The model-as-truth performance assessment is done on $T = 30$ (1990 - 2019 and 2030-2039) and a training period $T_1 = 20$ (1990 - 2019) as described in Section 2.6.2. The results are summarized in Figure 2.6 and Table 2.6 in form of median, 25% and 75% percentiles of the $N = 12$ (number of models) values.

Ensemble	B_T	B_{CM}	B_{IV}	$RMSE_{CM}$
E_{AVE}	-0.03	0.01	-0.42	0.96
E_{COE}	0.01	-0.01	-0.11	0.24
E_{MCE}	-0.02	-0.01	-0.36	0.50

Table 2.6: Model-as-truth performance comparison of different methods on NARCLiM data, median of trend bias (B_T), climatological monthly bias (B_{CM}), interannual variability bias (B_{IV}) and climatological monthly RMSE ($RMSE_{CM}$) on validation data.

As in the CMIP5 results (Figure 2.4 and Table 2.4) all the methods perform at the same level in B_T and B_{CM} metrics of the model-as-truth assessment. In B_{IV} and $RMSE_{CM}$ metrics E_{COE} performs better, while E_{AVE} performs worse than E_{MCE}.

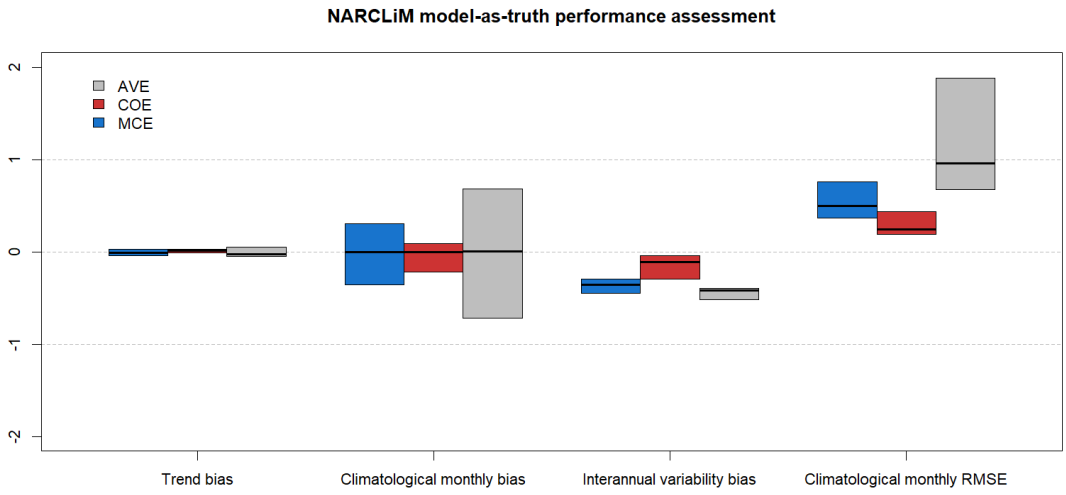

Figure 2.6: NARCLiM model-as-truth performance assessment results. Median, 25% and 75% percentiles of the $N = 12$ models.

2.3.3 KMA data

The KMA data is non-negative with a non-normal distribution of model outputs and observations as shown in panels a) and b) in Figure 2.7.

Applying the MCE method on the selected data with $T = 33$ (1973 - 2005) and a training period $T_1 = 22$ (1973 - 1994), we obtain a weighted ensemble mean E_{MCE} and compare it with outputs from other methods. As KMA data contains only summertime months, we analyse only its $RMSE_T$ and $RMSE_V$. We summarize KMA data properties together with the resulting ensemble's weights in Figure 2.7 and holdout cross-validation results in Table 2.7.

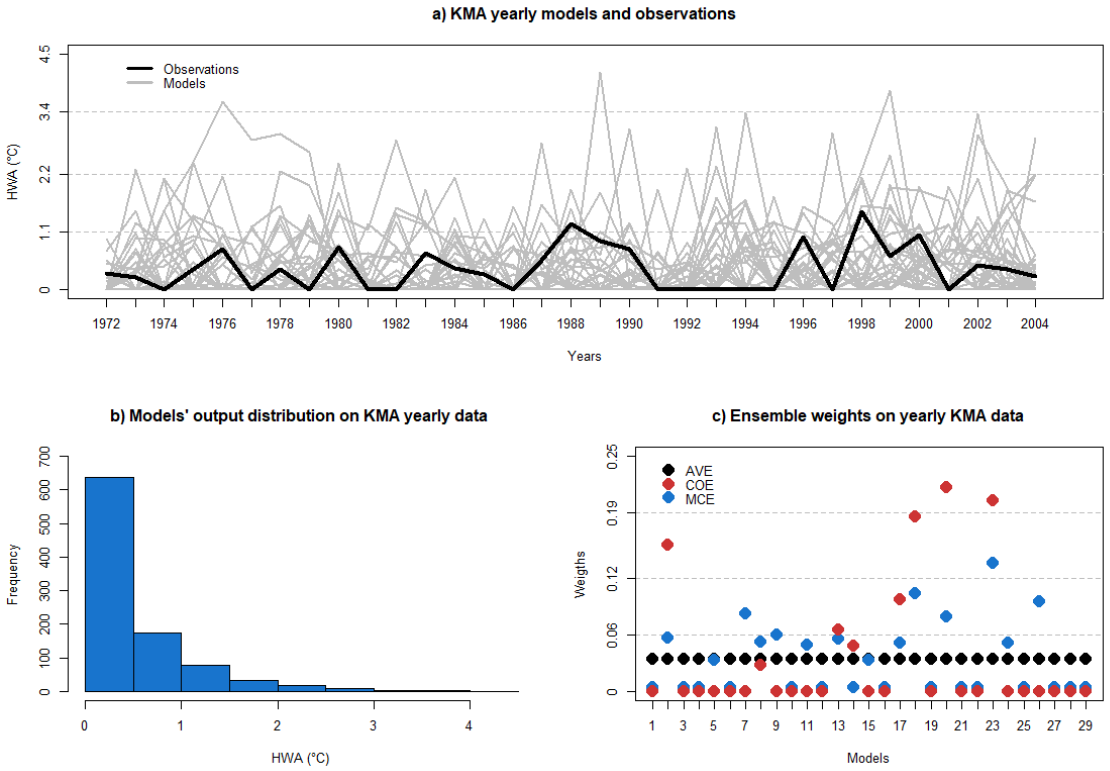

Figure 2.7: KMA data properties. **a)** Model outputs and observations. **b)** Model output distribution. **c)** AVE, COE and MCE weights.

$Ensemble$	$RMSE_T$	$RMSE_V$
E_{AVE}	0.36	0.5
E_{COE}	0.23	0.52
E_{MCE}	0.29	0.44

Table 2.7: Performance comparison of different methods on KMA data, RMSE on training ($RMSE_T$) and validation ($RMSE_V$) data.

We can see that MCE has the lowest RMSE and maintains the ensembles' diversity with a few models receiving zero weights. The COE method gives non-zero weights to only a small subset of models, which results in its performance on the validation period being lower compared to MCE.

2.4 Discussion

The obtained results indicate that Markov chains can be used to construct a better performing weighted ensemble mean with lower RMSE on validation data than commonly used methods like multi-model ensemble average and convex optimisation (Tables 2.3, 2.5 and 2.7). As the method's performance did not degrade from training to validation as much as COE, we are confident that it is less prone to over-fitting than linear optimisation methods. We attribute this advantage of the MCE method to its ability to maintain the ensemble's diversity while optimising its weights on the training period (Figures 2.3, 2.5 and 2.7), to mitigate model interdependence and to capture some of the nonlinear patterns in the data.

The MCE method also performs at the same level as other methods in terms of climatological metrics and model-as-truth performance assessment, which gives us confidence in its ability to be used for future estimation of climate variables.

However, as previous studies show (e.g. Masson and Knutti (2011); Sanderson et al. (2017)) and as discussed in Section 2.2.3, extending the MCE method to include spatial information would improve our ability to interpret the physical meaning of the resulting weights.

As the number of models increases, MCE tends to become closer to AVE weights (Figure 2.3), while being closer to COE with a smaller number of models (Figure 2.7). This phenomenon can be explained by a higher effect of diversity on performance in larger ensembles with normally distributed data (observations and model outputs) than in smaller ensembles like NARCLiM. The KMA data has an intermediate number of models and MCE produces a hybrid response which maintains ensemble diversity (a few models with zero weights) but does weight a small number of models more highly.

The MCE method is computationally cheap and is limited only by a software's ability to handle extreme numerical values. One limitation of the MCE method is its current inability to quantify the uncertainty of the resulting weighted ensemble mean. However,

we believe that given the stochastic nature of the method, this limitation can be overcome in future implementations. MCE performance can be further improved by combining it with other types of optimisation, e.g. linear. In addition, other nonlinear optimisation techniques, which would include more complex structures than simple Markov chains, can be developed based on our demonstrated results.

Finally, the MCE method doesn't require some of the assumptions necessary for the multi-model ensemble average method (e.g. models being reasonably independent and equally plausible as discussed by Knutti et al. (2017)) and it doesn't produce as many zero weights as the convex optimisation method, hence maintaining more of the models' diversity. We attribute the tendency of the COE method to set zero weights to some models to its property below:

Geometrically, the restrictions $w_j \geq 0, \sum_{j=1}^{N} w_j = 1$ describe a simplex in R^N that is a subset of the hyperplane with the equation $\sum_{j=1}^{N} w_j = 1$. Denote $w = (w_1, w_2, \ldots, w_N)$. The potential choice of weights that only satisfy the constraint $\sum_{j=1}^{N} w_j = 1$ without the non-negativity restriction represents any point in the hyperplane $P = \{w : \sum_{j=1}^{N} w_j = 1\}$. This hyperplane contains the simplex $S = \{w \in P : w_j \geq 0\}$. In general, the optimal point w^* for the unrestricted solution of the optimisation problem

$$\min_{w} \sum_{i=1}^{T} (\sum_{j=1}^{N} w_j M_{j,i} - O_i)^2, w \in P$$

will be outside the simplex. It is clear that the optimal point for the constrained solution on the simplex:

$$\min_{w} \sum_{i=1}^{T} (\sum_{j=1}^{N} w_j M_{j,i} - O_i)^2, w \in S$$

would be on the boundary of the simplex rather than in its interior. Indeed, if we assume that the optimal point for the constrained problem is certain $\tilde{w}$ in the interior of the simplex, we immediately arrive at a contradiction. Take then the point $\hat{w} = w^* + \lambda(\tilde{w} - w^*)$ with $\lambda \in (0, 1)$ chosen such that $\hat{w}$ is on the intersection of the line connecting w^* and $\tilde{w}$ with the boundary of the simplex. Because of the strict convexity of the function

$$f(w) = \sum_{i=1}^{T} (\sum_{j=1}^{N} w_j M_{j,i} - O_i)^2$$

we have:

$$f(\hat{w}) = f(w^* + \lambda(\tilde{w} - w^*)) = f(\lambda\tilde{w} + (1-\lambda)w^*) < \lambda f(\tilde{w}) + (1-\lambda)f(w^*) < f(\tilde{w})$$

in contradiction to the assumption that $\tilde{w}$ delivers the minimum over the simplex. Hence the optimisation on the simplex tends to deliver optimal points with some components equal to zero because they tend to be on the boundary of the simplex.

2.5 Conclusions

In this study, we presented a novel approach based on Markov chains to estimate model weights in constructing weighted climate model ensemble means. The complete MCE method was applied to selected climate datasets, and its performance was compared to two other common approaches (AVE and COE) using cross-validation holdout method and model-as-truth performance assessment with RMSE, trend bias, climatology monthly bias, interannual variability and climatological monthly RMSE metrics. The MCE method was discussed in detail, and its step-wise implementation, including mathematical background, was presented (Table 2.2).

The results of this study indicate that applying nonlinear ensemble weighting methods on climate datasets can improve future climate projection in terms of accuracy. Even a simple nonlinear structure such as Markov chains shows good performance on different commonly-used datasets compared to linear optimisation approaches. These results are supported by using standard performance metrics, cross-validation procedures and model-as-truth performance assessment. The developed MCE method is objective in terms of parameter selection, has a sound theoretical basis and has a relatively low number of limitations. It maintains ensemble diversity, mitigates model interdependence and captures some of the nonlinear patterns in the data while optimizing ensemble weights. It is also shown to perform well on non-Gaussian datasets. Based on the above, we are confident to suggest its application on other datasets and its usage for the future development of new nonlinear optimisation methods for weighting climate model ensembles.

Chapter 3

Varying weight MCE for spatially explicit climate data

3.1 Introduction

As discussed in Chapter 2 (weighted) means of multi-model ensembles are generally better correlated with observations than a single climate model. How to best select weights for a weighted multi-model ensemble remains an open research question. In this chapter we extend the Markov chain ensemble (MCE) method presented in Chapter 2 to illustrate the potential of applying nonlinear ensemble weighting methods. We call this extended method a varying weight Markov chain ensemble (VMCE) method with a variable weighting scheme based on monthly time series.

We apply VMCE and other methods described below on spatially explicit data from the sixth Coupled Model Intercomparison Project (CMIP6), and analyse and compare their performance in different regions. By applying the methods to a large number of cells as opposed to one global time series as in Chapter 2, we investigate the methods' performance across different regions. To accomplish this task we use data from CMIP6 which contains the latest update on expected future climate change based on a new generation of climate

models (Eyring et al. (2016)). The CMIP6 data is shown to differ from CMIP5 (Chao-An et al. (2021), Bourdeau-Goulet et al. (2021), etc.) and provides complimentary empirical evidence of our method's performance in addition to the results obtained and analysed in Chapter 2.

To further diversify scenario for application of VMCE method, we use climate variables with Gaussian and non-Gaussian distributions and different scales (non-negative and both positive and negative). To understand the advantages and limitations of VMCE method we analyse the patterns in VMCE weights and in the corresponding spatial climate data in different climate zones and geographical regions.

The performance of VMCE is compared to the performance of other methods discussed earlier (AVE, COE, and MCE) and to the introduced VCOE method based on the same varying weights scheme as VMCE. We analyse if the varying weights scheme application is beneficial for methods other than MCE and to what extent.

In what follows, we will first introduce the datasets used in this chapter in section 3.2.1, then outline the proposed varying weights scheme VMCE and its implementation, and an extension of the COE method proposed by Bishop and Abramowitz (2013), which we call VCOE, which allows for varying weights under the COE framework, performance metrics and cross-validation procedures are described in Section 3.2.2. We will compare the proposed (VMCE) method's performance to the commonly used multi-model ensemble average (AVE) method (Lambert and Boer (2001)), the convex optimisation (COE) method, its modification with varying weights scheme (VCOE) and Markov chain ensemble method (MCE) discussed in Chapter 2. We will present the results in Section 3.3, followed by a discussion in Section 3.4 and conclusion in Section 3.5.

3.2 Methods

3.2.1 Data

Here we first describe the datasets used in this study. We have chosen publicly available datasets from CMIP6, the Berkeley Earth land temperature record and Global Precipitation Climatology Centre (GPCC). All the datasets are spatially explicit and are converted to the same resolution of 2.5° x 2.5°.

CMIP6 model data: The climate models' output data is obtained from https://climexp.knmi.nl with 35 models (one ensemble member per model, native model resolution, all models available on 26/10/2021, see Table 3.1) and the period of 1901 – 2100 is selected for the analysis. The models contain temperature anomalies (monthly averages) compared to the reference period of 1961-1990 (Taylor et al. (2011)) and precipitation (monthly averages) data. This dataset contains several clusters of dependent models (see Figures 3.1 and 3.2 below) with temperature having Pearson correlation coefficient averaged over land points between 0.85 and 1, and precipitation - between 0.2 and 0.5. Temperature data has both positive and negative data values and a relatively low variability (see Figure 3.3). Precipitation data has non-negative values, a high variability and a non-Gaussian distribution (see Figure 3.4).

Model number	Model name	Model number	Model name
1	ACCESS1-0	19	GISS-E2-H
2	ACCESS1.3	20	GISS-E2-R
3	BCC-CSM1-1	21	HadGEM2-AO
4	BNU-ESM	22	HadGEM2-CC
5	CanESM2	23	HadGEM2-ES
6	CCSM4	24	INMCM4
7	CESM1-BGC	25	IPSL-CM5A-LR
8	CESM1-CAM5	26	IPSL-CM5A-MR
9	CMCC-CM	27	IPSL-CM5B-LR
10	CMCC-CMS	28	MIROC5
11	CNRM-CM5	29	MIROC-ESM
12	CSIRO-Mk3-6-0	30	MIROC-ESM-CHEM
13	EC-EARTH	31	MPI-ESM-LR
14	FGOALS-G2	32	MPI-ESM-MR
15	FIO-ESM	33	MRI-CGCM3
16	GFDL-CM3	34	NorESM1-M
17	GFDL-ESM2G	35	NorESM1-ME
18	GFDL-ESM2M		

Table 3.1: CMIP6 models used in this study. Assigned model numbers with respective original model names.

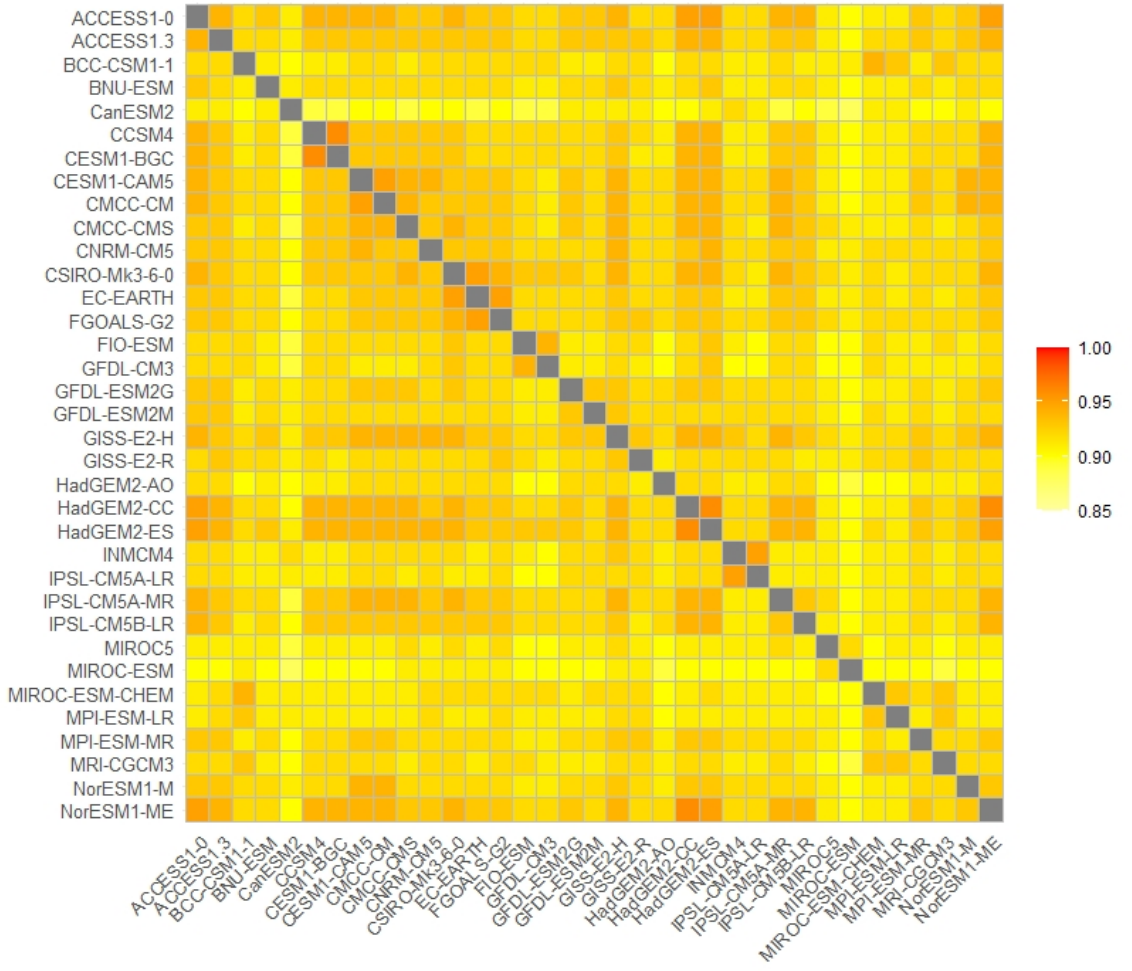

Figure 3.1: CMIP6 model outputs' correlation matrix for temperature data with Pearson correlation coefficient averaged over land points between 0.85 and 1. Dark colour regions show clusters of highly dependent models.

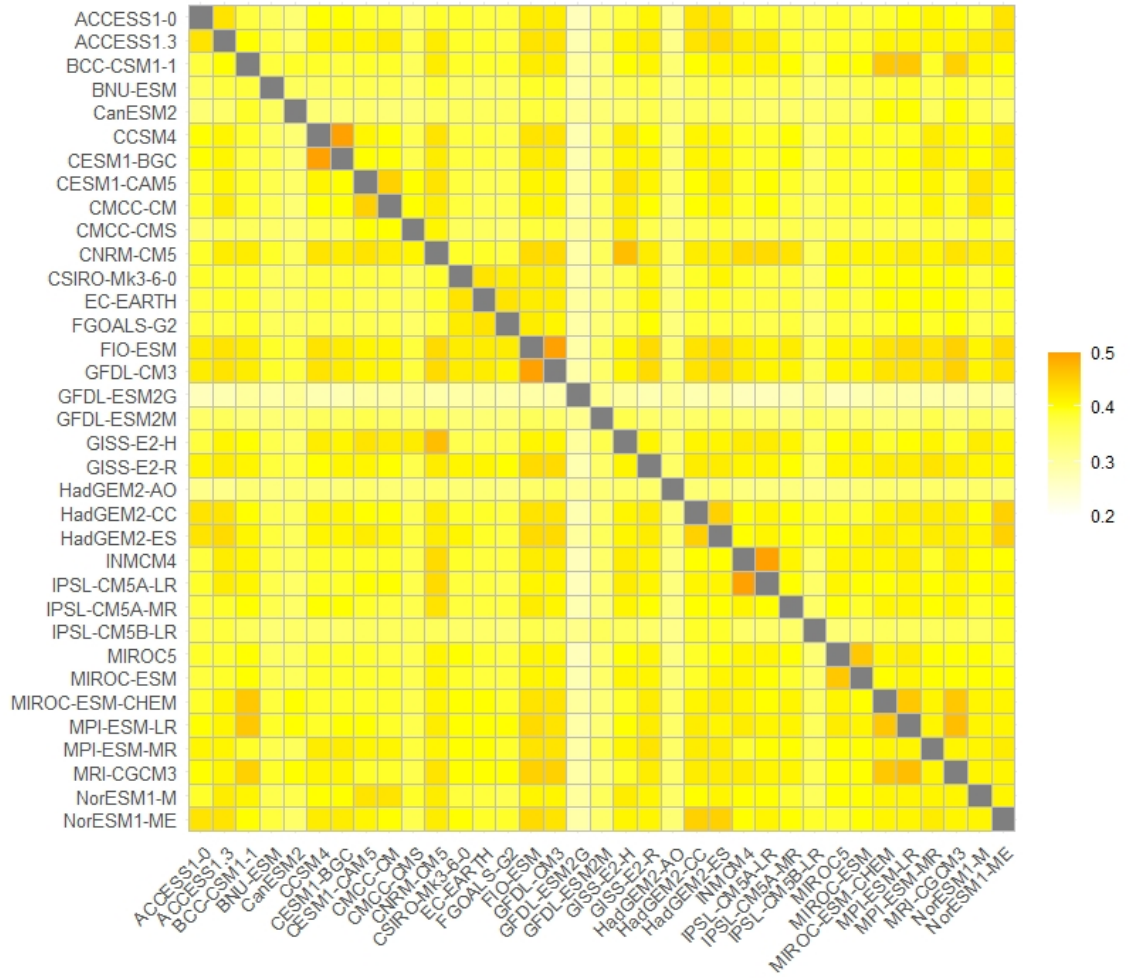

Figure 3.2: CMIP6 model outputs' correlation matrix for precipitation data with Pearson correlation coefficient averaged over land points between 0.2 and 0.5. Dark colour regions show clusters of highly dependent models.

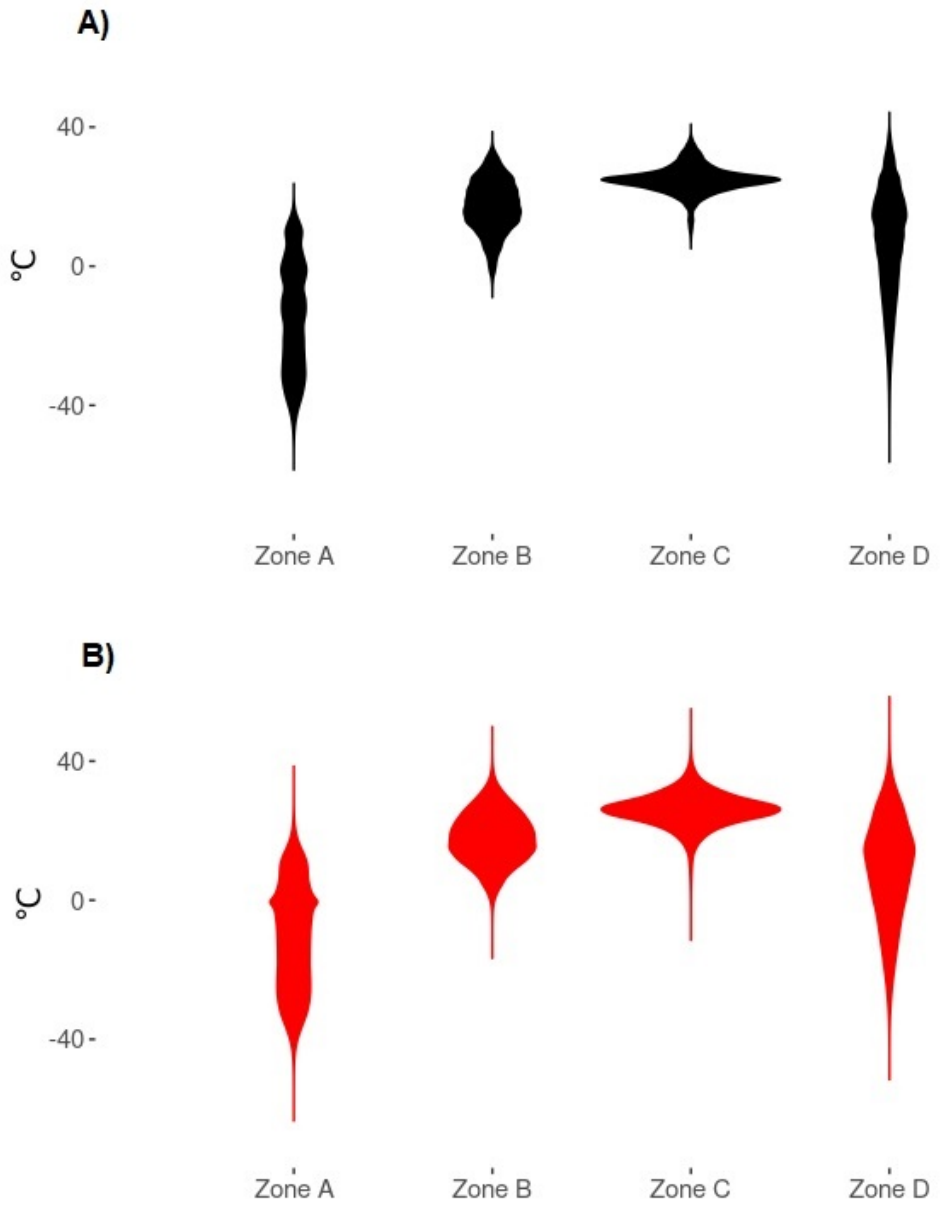

Figure 3.3: Violin plots showing distribution of CMIP6 temperature data over all land points. Zone A includes all land points in polar regions (southern and northern) combined (with latitude south from 66.5°S or north from 66.5°N). Zone B includes all land points in southern regions between polar and tropical (with latitude between 66.5°S and 23.5°S). Zone C includes all land points in tropical regions (with latitude between 23.5°S and 23.5°N). Zone D includes all land points in Northern regions between tropical and polar (with latitude between 23.5°N and 66.5°N). A) Violin plots showing distribution of observations. B) Violin plots showing distribution of all model outputs (See Table 3.1)

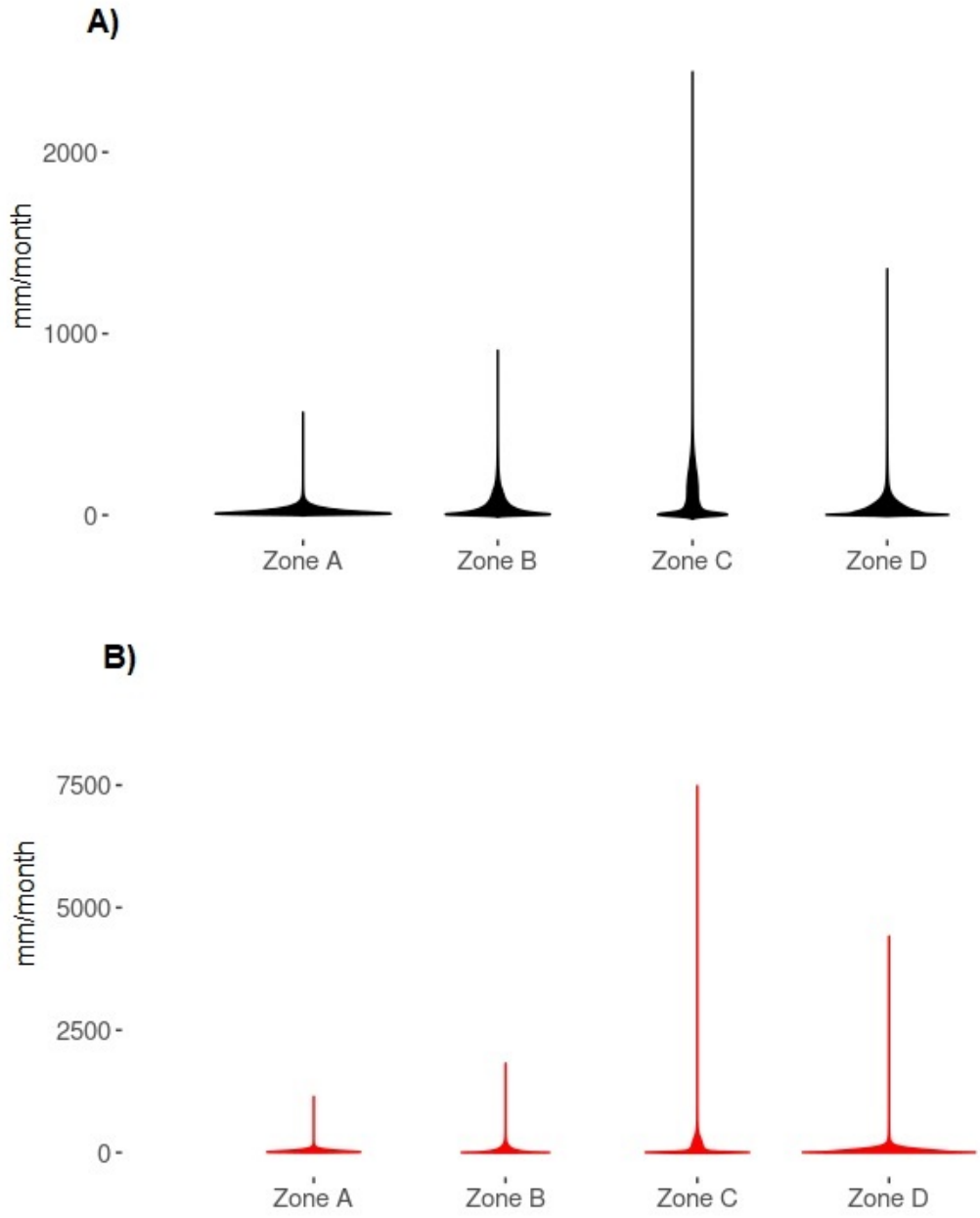

Figure 3.4: Violin plots showing distribution of CMIP6 precipitation data over all land points. Zone A includes all land points in polar regions (southern and northern) combined (with latitude south from 66.5°S or north from 66.5°N). Zone B includes all land points in southern regions between polar and tropical (with latitude between 66.5°S and 23.5°S). Zone C includes all land points in tropical regions (with latitude between 23.5°S and 23.5°N). Zone D includes all land points in Northern regions between tropical and polar (with latitude between 23.5°N and 66.5°N). A) Violin plots showing distribution of observations. B) Violin plots showing distribution of all model outputs (See Table 3.1)

Berkley Earth data: The temperature observational data is obtained from http://berkeleyearth.org/data/ with 1° x 1° degree resolution and the period of 1901-2020 is selected for the analysis. The properties of this data are summarised in Figure 3.5 below.

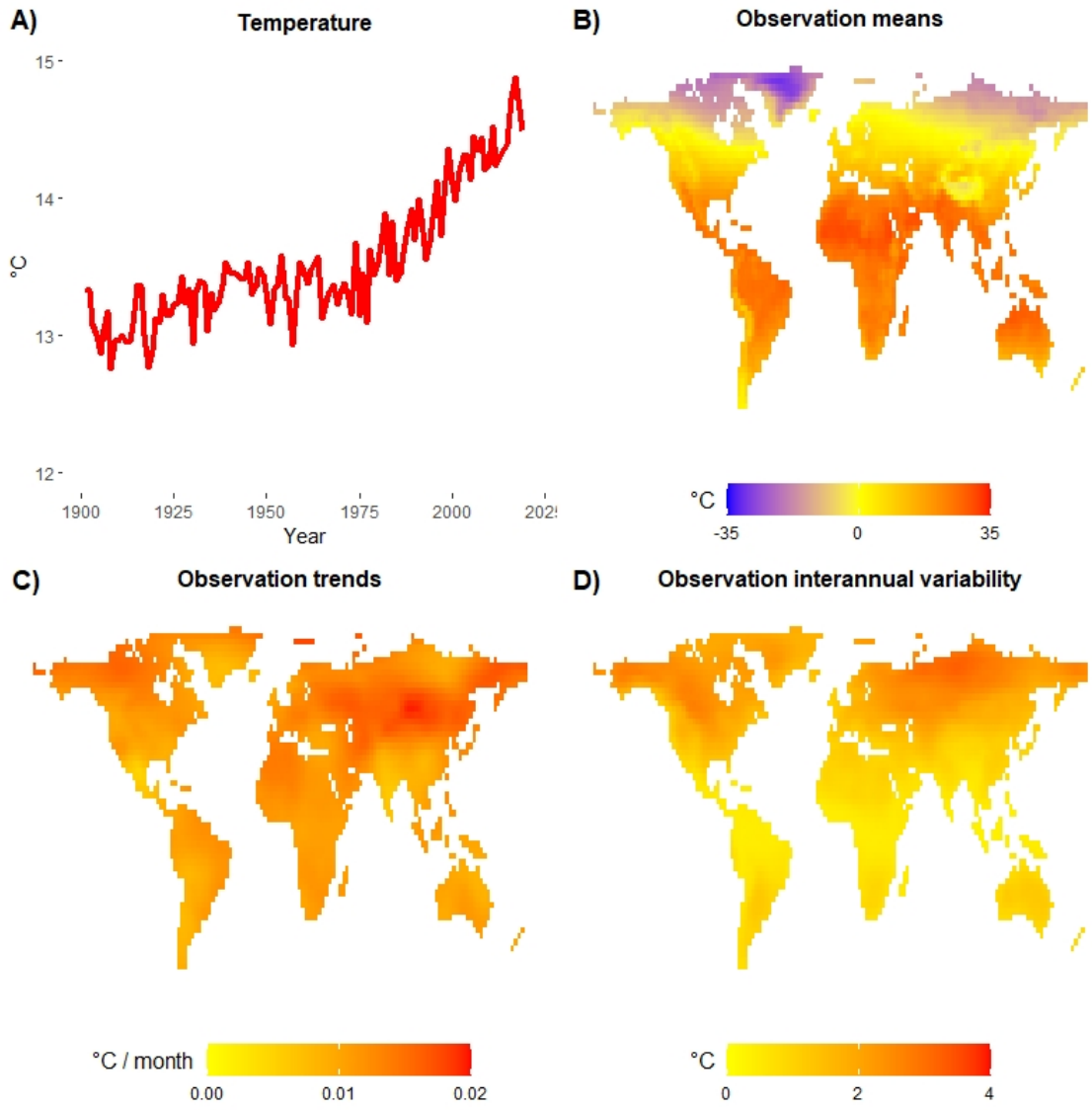

Figure 3.5: Properties of temperature observational data (Berkley Earth). **A)** Global yearly time series. **B)** Geo-spatial distribution of observation time series' means. **C)** Geo-spatial distribution of observation time series' trends. **D)** Geo-spatial distribution of observation time series' interannual variability.

GPCC data: The precipitation observational data is obtained from https://climexp.knmi.nl with 1° x 1° degree resolution and the period of 1901-2020 is selected for the analysis. The properties of this data are summarised in Figure 3.6 below.

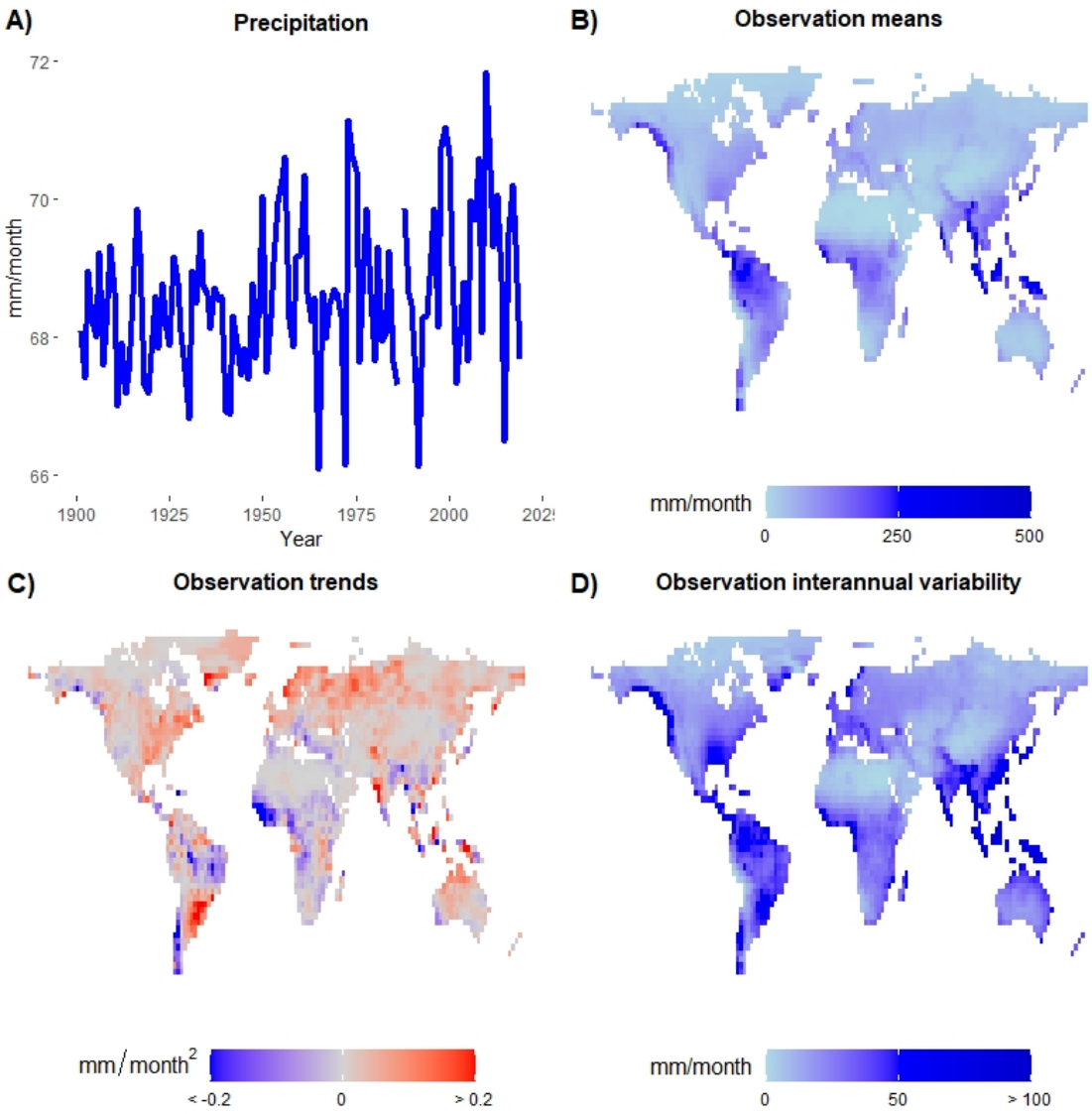

Figure 3.6: Properties of precipitation observational data (GPCC). **A)** Global yearly time series. **B)** Geo-spatial distribution of observation time series' means. **C)** Geo-spatial distribution of observation time series' trends. **D)** Geo-spatial distribution of observation time series' interannual variability.

3.2.2 Methods

The Markov chain ensemble (MCE) method is described in detail in Chapter 2 and here we extend it by pre-processing input data and post-processing output data to have a varying set of weights for each month. We call this method a Varying Markov Chain Ensemble (VMCE) method and implement it as (Table 3.2):

Input:

- length of training period T_1, and
- historical observations O_t, at times $t = 1, \ldots, T_1$, and
- climate model output $M_{i,t}$, at times $t = 1, \ldots, T_1$, for $i = 1, \ldots, N$ models, and
- an initialised number of simulations L
- an initialised σ interval $[\sigma_{min}, \sigma_{max}]$
- an initialised transition matrix P^0 of N x N size

Step 1. For each month m, construct a subset of historical observations $O^m = (O_m, O_{m+12}, O_{m+24}, ...)$.

Step 2. Construct the corresponding subset of model outputs $M_i^m = (M_{i,m}, M_{i,m+12}, M_{i,m+24}, ...)$ for $i = 1, ..., N$ models.

Step 3. Apply the MCE method according to (Table 2.2) with O^m and M^m as inputs and store the result as MCE^m.

Step 4. Repeat Steps 1-3 for each month $m = 1, ..., 12$

Step 5. Construct the resulting $VMCE$ by combining all MCE^m data as $VMCE = MCE_1^1, MCE_1^2, MCE_1^3, ..., MCE_1^{12}, MCE_2^1, MCE_2^2, MCE_2^3, ...$

Table 3.2: The varying weight Markov Chain ensemble (VMCE) algorithm.

The VMCE algorithm can be applied on global and local time series. In this study we apply it (and other methods we compare it with) on all land points of a 144 by 72 global grid. The main novelty of VMCE (compared to MCE and other methods) is utilisation of the repeating monthly pattern for optimisation purposes. Intuitively, this approach optimises

the weighted ensemble mean to represent year-to-year climate change more efficiently than the approaches that have to represent month-to-month variation as well. As the VMCE outcome consists of twelve sets of weights instead of a one set it gives the optimisation process more degrees of freedom. We demonstrate that those additional degrees of freedom are beneficial for optimisation in terms of climatological metrics we will describe below. We will also demonstrate that those additional degrees of freedom do not cause over-fitting to training data by using cross-validation procedures.

Since the ideas behind VMCE can also be applied to the COE method of Bishop and Abramowitz (2013), here we provide an extension to the standard COE method, which we refer to as the Varying Convex Optimisation Ensemble (VCOE) method. Table 3.3 provides the detailed description of the algorithm (which is also applied on all land points of a 144 by 72 global grid in this study):

Input:

- length of training period T_1, and
- historical observations O_t, at times $t = 1, \ldots, T_1$, and
- climate model output $M_{i,t}$, at times $t = 1, \ldots, T_1$, $i = 1, \ldots, N$

Step 1. For each month m, construct a subset of historical observations $O^m = (O_m, O_{m+12}, O_{m+24}, ...)$.

Step 2. Construct the corresponding subset of model outputs $M_i^m = (M_{i,m}, M_{i,m+12}, M_{i,m+24}, ...)$ for $i = 1, ..., N$ models.

Step 3. Apply the COE method (described in Section 2.2.4) with O^m and M^m as inputs and store the result as COE^m.

Step 4. Repeat Steps 1-3 for each month $m = 1, ..., 12$.

Step 5. Construct the resulting $VCOE$ by combining all COE^m data as $VCOE = COE_1^1, COE_1^2, COE_1^3, ..., COE_1^{12}, COE_2^1, COE_2^2, COE_2^3, ...$

Table 3.3: The varying weight convex optimization ensemble (VCOE) algorithm.

3.2.3 Parameter settings

We use the same parameters ($L = 3000, \sigma \in [0.1, 1]$) for VMCE method as for MCE (see Section 2.2.2.1 Parameter sensitivity). For COE method the same default settings as in Chapter 2 are used (see Fu et al. (2020)).

3.2.4 Performance metrics

Here we describe a set of metrics which we will use to compare the performance of our method. These include root mean squared error ($RMSE$), climatological monthly RMSE ($RMSE_{CM}$), climatology monthly bias (B_{CM}), trend bias (B_T) and interannual variability (B_{IV}). The performance metrics are applied on AVE, COE, MCE, VCOE and VMCE ensembles weighted means which are calculated according to equation 3.1 below.

$$E_t = \sum_{m=1}^{12} \sum_{y=1}^{Y} \sum_{j=1}^{N} w_{j,m} M_{j,12(y-1)+m} \tag{3.1}$$

with $w_{j,m}$ being a j model weight for month m and $\sum_{j=1}^{N} w_{j,m} = 1$, $w_{j,m} > 0$ for $j = 1, \ldots, N$ and $m = 1, \ldots, 12$. Y is the total number of years and $M_{j,12(y-1)+m}$ denotes the value of model j in year y, month m.

The root mean squared error ($RMSE$) was described in detail in Chapter 2 (Section 2.2.5.1). It is a common way to measure the error of a mathematical model in predicting quantitative data. Here it is used to measure the differences between the weighted ensemble means and observations. It is a metric that COE, MCE, VCOE and VMCE are optimized for on training periods. $RMSE$ is calculated using equation 3.2 below.

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (E_t - O_t)^2}, \tag{3.2}$$

with E_t being a weighted ensemble mean calculated according to equation 3.1. T is the total number of time steps, $M_{j,t}$ denotes the value of model j at time step t and O_t is the observed value at point t.

Climatological monthly RMSE ($RMSE_{CM}$) is calculated according to Equation 3.3 on climatological monthly means of weighted ensemble values and observations on validation data. We utilise this metric to evaluate the average ensemble monthly prediction bias.

$$RMSE_{CM} = \sqrt{\frac{1}{12}\sum_{m=1}^{12}\left(\frac{\sum_{y=1}^{Y} E_{12(y-1)+m} - \sum_{y=1}^{Y} O_{12(y-1)+m}}{Y}\right)^2}, \tag{3.3}$$

with $E_{12(y-1)+m}$ being a weighted ensemble mean calculated according to equation 3.1 for year $y = 1, \ldots, Y$, month $m = 1, \ldots, 12$. Y is the total number of years and $O_{12(y-1)+m}$ is the observed value in year y, month m.

Climatology monthly bias (B_{CM}) is calculated using equation 3.4 as mean of the differences between the mean of the weighted ensemble and the observation for each month. We utilise this metric to evaluate the average absolute ensemble monthly bias.

$$B_{CM} = \sum_{m=1}^{12} \frac{|\sum_{y=1}^{Y} E_{12(y-1)+m} - \sum_{y=1}^{Y} O_{12(y-1)+m}|}{Y} \tag{3.4}$$

with $E_{12(y-1)+m}$ being a weighted ensemble mean calculated according to equation 3.1 for year $y = 1, \ldots, Y$, month $m = 1, \ldots, 12$. Y is the total number of years and $O_{12(y-1)+m}$ is the observed value in year y, month m. We utilise this metric to evaluate the absolute ensemble monthly trend bias.

Trend bias (B_T) is calculated using equation 3.5 as a mean of the differences between the inclination parameters α^E in weighted ensembles and the inclination parameters α^O observations estimated using linear functions $y = \alpha x + \beta$ for each month.

$$B_T = \sum_{m=1}^{12} |\alpha_m^E - \alpha_m^O| \tag{3.5}$$

with $E_{12(y-1)+m} \sim \alpha^E(12(y-1)+m) + \beta^E$ being a linear approximation of the ensemble and $O_{12(y-1)+m} \sim \alpha^O(12(y-1)+m)+\beta^O$ being a linear approximation of the observations for each year $y = 1, \ldots, Y$ and each month $m = 1, \ldots, 12$. Y is the total number of years and $O_{j,12(y-1)+m}$ is the observed value in year y, month m.

Interannual variability (B_{IV}) is calculated using equation 3.6 as the mean of differences between the standard deviation of detrended weighted ensemble and the standard deviations of detrended observations. We utilise this metric to evaluate the absolute ensemble interannual variability bias.

$$B_{IV} = \sum_{m=1}^{12} |\sigma_m^E - \sigma_m^O| \tag{3.6}$$

with ensemble standard deviation σ_m^E and observations standard deviation σ_m^O calculated according to formulas 3.7 and 3.8 below.

$$\sigma_m^E = \sqrt{\frac{\sum_{y=1}^{Y}(E^*_{12(y-1)+m} - \frac{\sum_{y=1}^{Y} E^*_{12(y-1)+m}}{Y})^2}{Y}} \tag{3.7}$$

with $E^*_{12(y-1)+m} = E_{12(y-1)+m} - (\alpha_m^E(12(y-1)+m) + \beta_m^E)$ being detrended weighted ensemble values for each year $y = 1, \ldots, Y$ and each month $m = 1, \ldots, 12$. a^E is inclination parameter in linear approximation $E_{12(y-1)+m} \sim \alpha^E(12(y-1)+m) + \beta^E$ of the weighted ensemble for each year $y = 1, \ldots, Y$ and each month $m = 1, \ldots, 12$. Y denotes the total number of years.

$$\sigma_m^O = \sqrt{\frac{\sum_{y=1}^{Y}(O^*_{12(y-1)+m} - \frac{\sum_{y=1}^{Y} O^*_{12(y-1)+m}}{Y})^2}{Y}} \tag{3.8}$$

with $O^*_{12(y-1)+m} = O_{12(y-1)+m} - (\alpha^O_m(12(y-1)+m) + \beta^O_m)$ being detrended observations for each year $y = 1, \ldots, Y$ and each month $m = 1, \ldots, 12$. a^O is inclination parameter in linear approximation $O_{12(y-1)+m} \sim \alpha^O(12(y-1)+m) + \beta^O$ of the observations for each year $y = 1, \ldots, Y$ and each month $m = 1, \ldots, 12$. Y is the total number of years. $O_{j,12(y-1)+m}$ denotes the observed value in year y, month m.

3.2.5 Cross-validation procedures

Here we describe two cross-validation procedures we use in this study to assess how the results obtained on calibration data will perform on new independent datasets. For the time period where observational data is available we employ a hold out method where part of the observational data is used for training and another part is used for validation. For the time period where observational data is not available we employ a model-as-truth performance assessment where each of the models serves as a proxy for the missing observational data.

Holdout method is described in detail in Chapter 2 (Section 2.2.6.1) and is based on splitting the dataset into training and validation sets. The goal of cross-validation is to examine the model's ability to predict new data that was not used in estimating the required parameters.

Model-as-truth performance assessment is described in detail in Chapter 2 (Section 2.2.6.2) and is based on selecting one model as a true model (pseudo-observations) with the remaining models used to build a weighted ensemble mean that best estimates the true model over the historical period. This weighted ensemble mean is then tested against the future projections of the true model. This procedure is repeated N times with each of the ensemble members being a true model where N is the total number of the ensemble members.

3.3 Results

3.3.1 Temperature data

We present the summary of the method comparison analysis for temperature in Tables 3.4 and 3.5 below. The values are the weighted means of all land points on 144 by 72 global grid with their respective area sizes as weights.

Ensemble	$RMSE$	$RMSE_{CM}$	B_{CM}	B_T	B_{IV}
AVE	2.59	2.07	1.84	0.01	0.95
COE	1.84	1.09	0.95	0.01	0.79
MCE	2.03	1.40	1.20	0.01	0.91
VCOE	**1.40**	0.43	0.31	0.01	**0.70**
VMCE	1.43	**0.37**	**0.21**	0.01	0.86

Table 3.4: Average temperature results using all land points weighted according to their area sizes on training period **(years 1901-1980)**. The minimum values in each column are emphasised in bold.

The VMCE method is capable of finding optimal weights for $RMSE_{CM}$ and B_{CM} metrics, while the VCOE method performs best for $RMSE$ and B_{IV} metrics. As discussed in Chapter 2 convex optimisation is generally capable of finding the best weights for $RMSE$ on training period. However, it does not produce the best performance for the same metrics on validation period as demonstrated below in Table 3.5. VCOE method has the lowest interannual variability bias (B_{IV}) as convex optimisation tends to exclude part of the models from the resulting ensemble by assigning 0 weights (see Figures 2.3,2.5 and 2.7) which brings its interannual variability closer to the internal variability of a single model and observations (as a single model can be seen as a representation of observations). We validate these findings on training period by examining the results on validation period summarised in Table 3.5.

Ensemble	$RMSE$	$RMSE_{CM}$	B_{CM}	B_T	B_{IV}
AVE	2.62	2.11	1.86	0.02	0.94
COE	1.91	1.18	1.02	0.02	0.78
MCE	2.10	1.48	1.27	0.02	0.91
VCOE	1.63	0.66	0.53	0.02	**0.67**
VMCE	**1.58**	**0.64**	**0.47**	0.02	0.85

Table 3.5: Average temperature results using all land points weighted according to their area sizes on validation period **(years 1981-2020)**. The minimum values in each column are emphasised in bold.

The VMCE method outperforms all other methods in terms of $RMSE$, $RMSE_{CM}$ and B_{CM} on valdiation period. In addition, VMCE has better $RMSE$ performance than VCOE on validation data, while having worse performance on training data. This indicates that VMCE is less prone to over-fitting the training data (as discussed in Chapter 2). Together with the results on the training period, validation period results indicate that the trend bias B_T performance level is the same for all the methods. The high performance of VCOE and VMCE methods compared to COE and MCE respectively indicates that varying sets of weights can be applied to different methods and not only to MCE to potentially increase their performance.

The detailed results for each climatological metric are presented as maps and distributions of the results for different methods on training and validation periods together with model-as-truth experiments' results (see Figures 3.7, 3.8, 3.9 for $RMSE$, Figures 3.10, 3.11, 3.12 for $RMSE_{CM}$, Figures 3.13, 3.14, 3.15 for B_{CM}, Figures 3.16, 3.17, 3.18 for B_T and Figures 3.19, 3.20, 3.21 for B_{IV}). The violin plots describe the distribution of metric values for each method using all land points with white lines showing median values. The violin plots title describes the metric displayed on violin and spatial plots in form of $[Metricabbreviation]^{[Time\ period]}(Temperature)$ with metric abbreviation being $RMSE$ for root mean squared error, $RMSE_{CM}$ for climatological monthly RMSE, B_{CM} for climatological monthly bias, B_T for trend bias and B_{IV} for interannual variability; time period being Tr for training period (years 1901-1980), V for validation period (years 1981-2020) and $M-a-T$ for model-as-truth experiment period (years 2021-2100). The spatial plots

describe the spatial distribution of metric values for each method with according method names in title. All the plots exclude small portions of data which constitute long tails of the data distributions. The maximum included values are specified in violin plots' axis and in the captions. Darker colours indicate geographical locations with higher metric values.

We analyse the $RMSE$ performance of all the methods on training period (Figure 3.7), validation period (Figure 3.8) and in model-as-truth experiment (Figure 3.9) below.

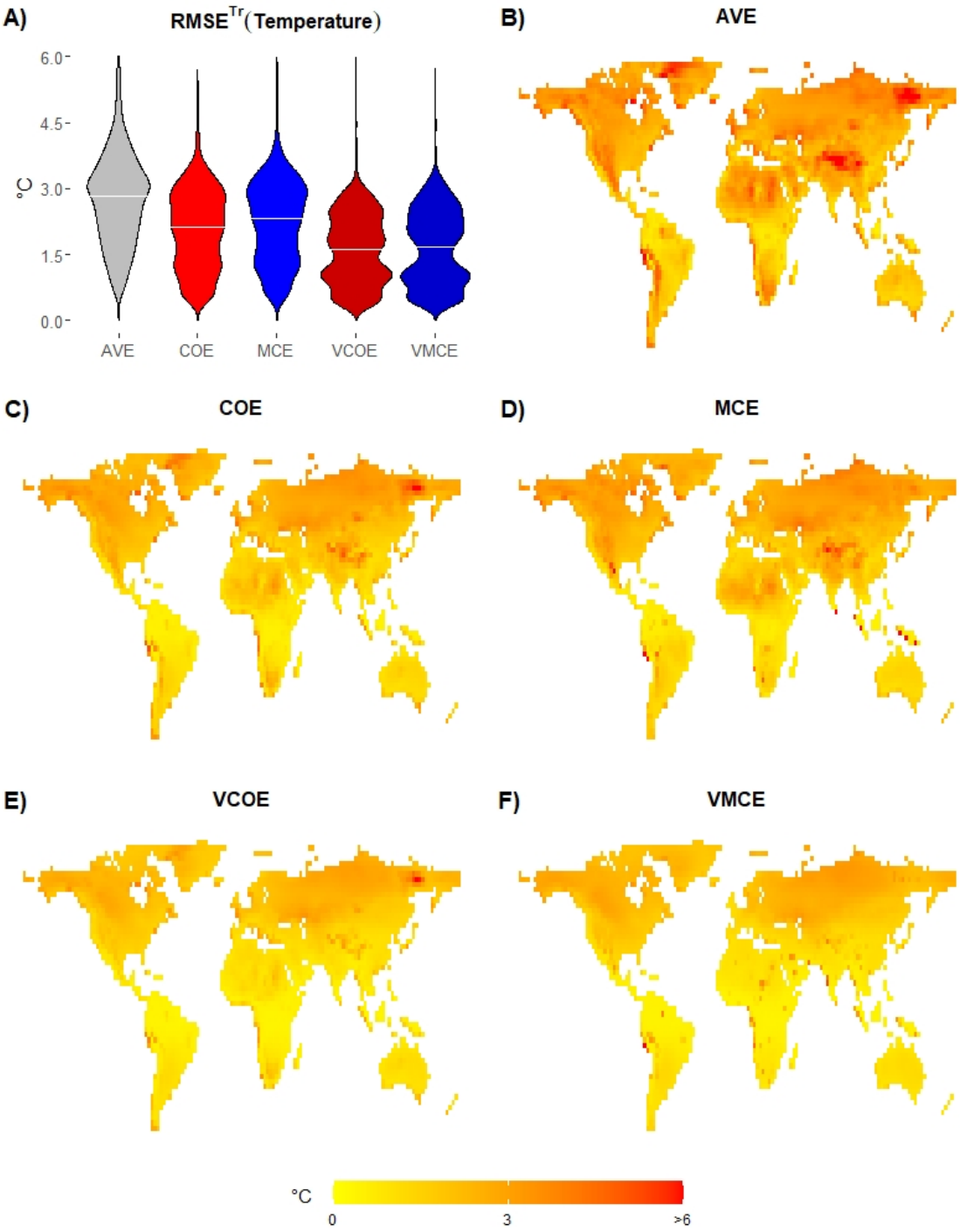

Figure 3.7: Root mean squared error ($RMSE$) results on training period **(years 1901-1980)** for temperature. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 6.0. **B) - F)** Geospatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

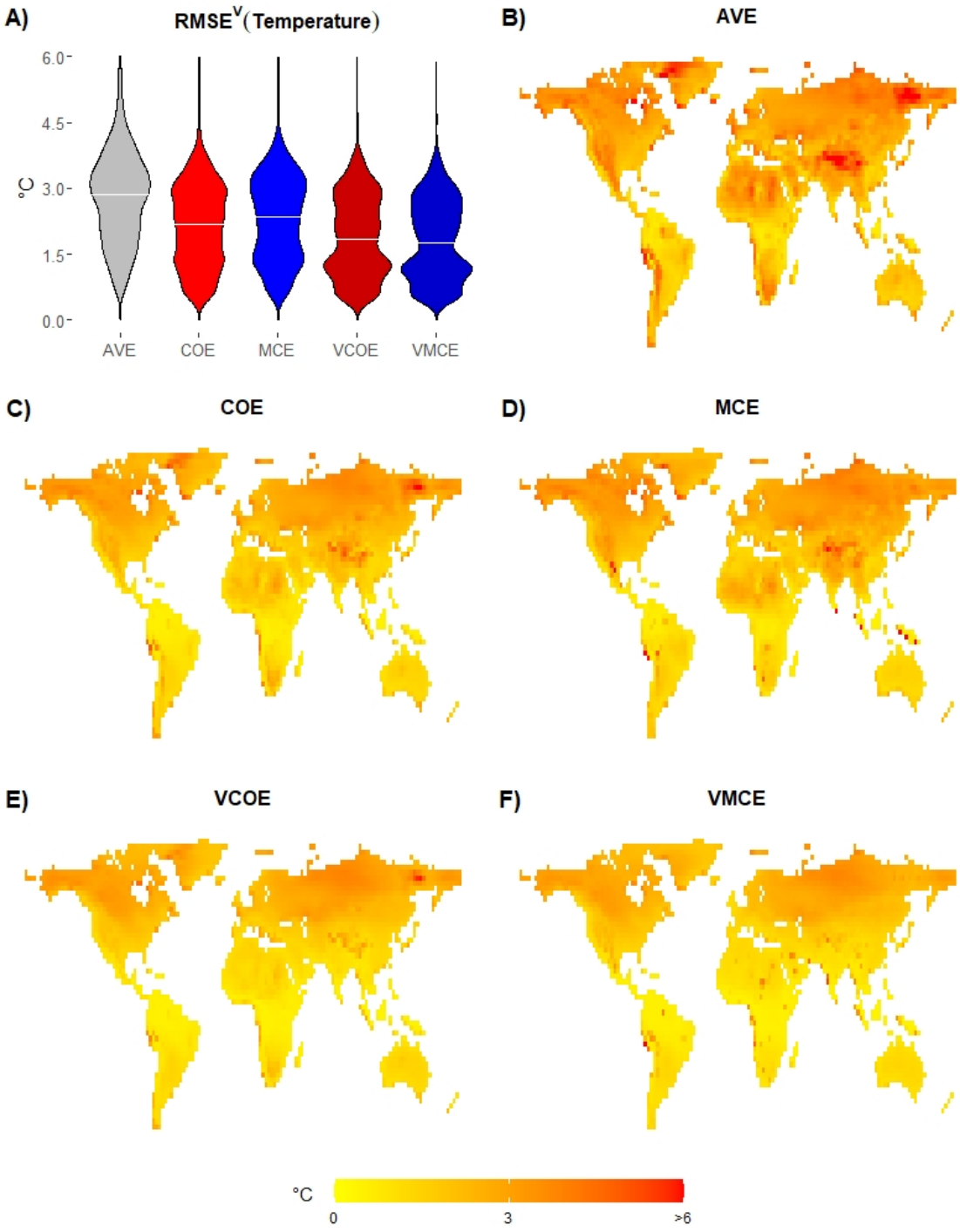

Figure 3.8: Root mean squared error ($RMSE$) results on validation period **(years 1981-2020)** for temperature. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 6.0. **B) - F)** Geospatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

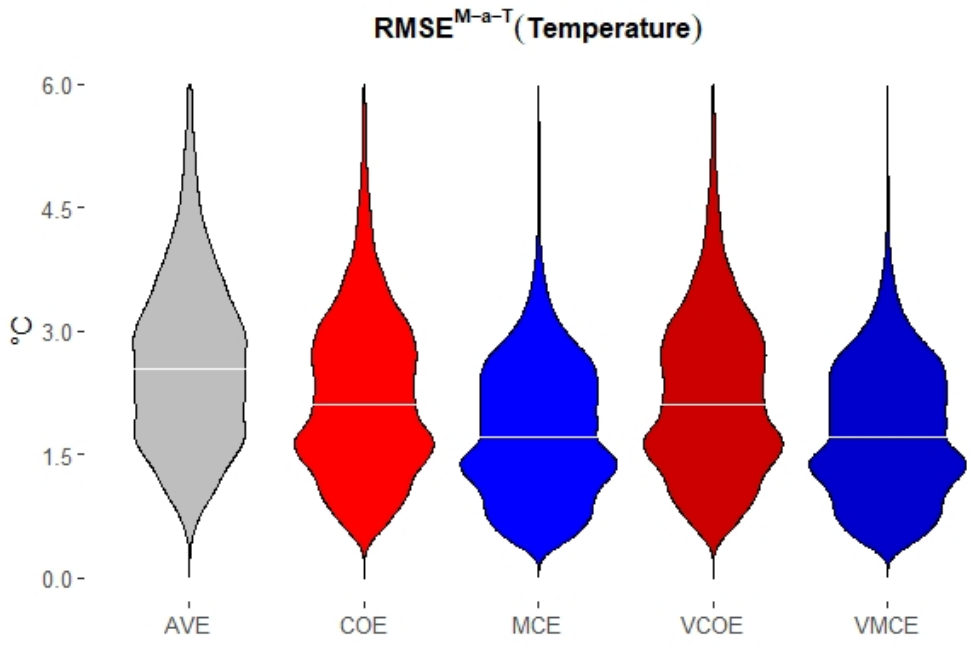

Figure 3.9: Violin plot showing model-as-truth experiment root mean squared error ($RMSE$) results for temperature during years 2021-2100 using all land points with white lines showing median values. The Y-axis is cut at 6.0.

The VMCE method has the best $RMSE$ performance on validation period and a relatively low variance of the results. This confirms that VMCE is able to find optimal weights in terms of $RMSE$ and is more flexible than any of the linear methods due to its higher degrees of freedom. The lower spread indicates that the VMCE results are more consistent (compared to competing methods) across the spatial grid with the exception of a few grid cells visible on both Figures 3.7 and 3.8 located mostly in desert areas. Model-as-truth experiment results indicate that MCE and VMCE methods perform better in terms of $RMSE$ than other methods and have lower spreads of the results. As VMCE has similarly good $RMSE$ performance on validation period as well as in model-as-truth experiment it indicates that VMCE is applicable for future temperature estimation in terms of minimising $RMSE$. We analyse the $RMSE_{CM}$ performance of all the methods on training (Figure 3.10) and validation (Figure 3.11) periods and in model-as-truth experiment (Figure 3.12).

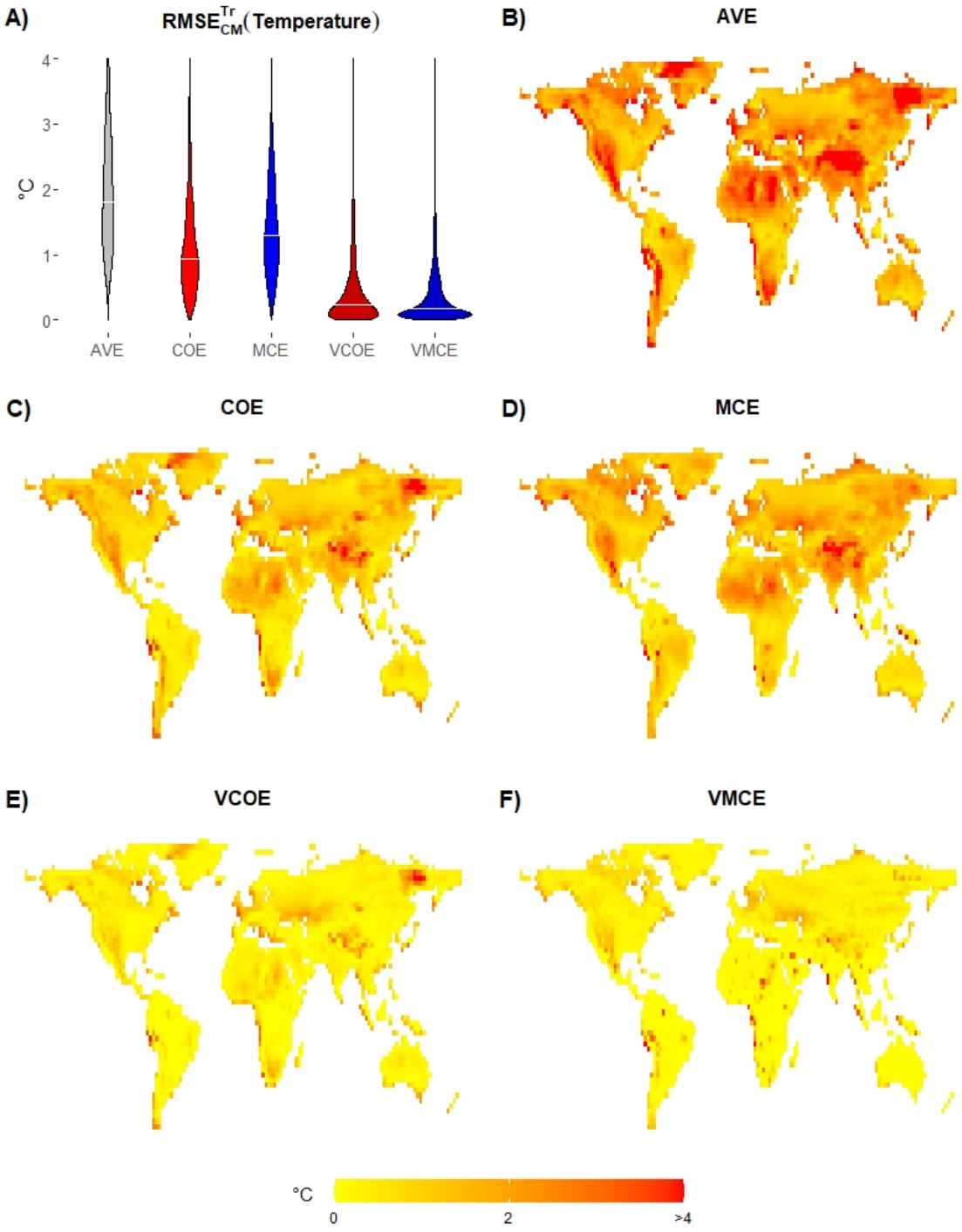

Figure 3.10: Climatological monthly RMSE ($RMSE_{CM}$) results on training period **(years 1901-1980)** for temperature. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 4.0. **B) - F)** Geo-spatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

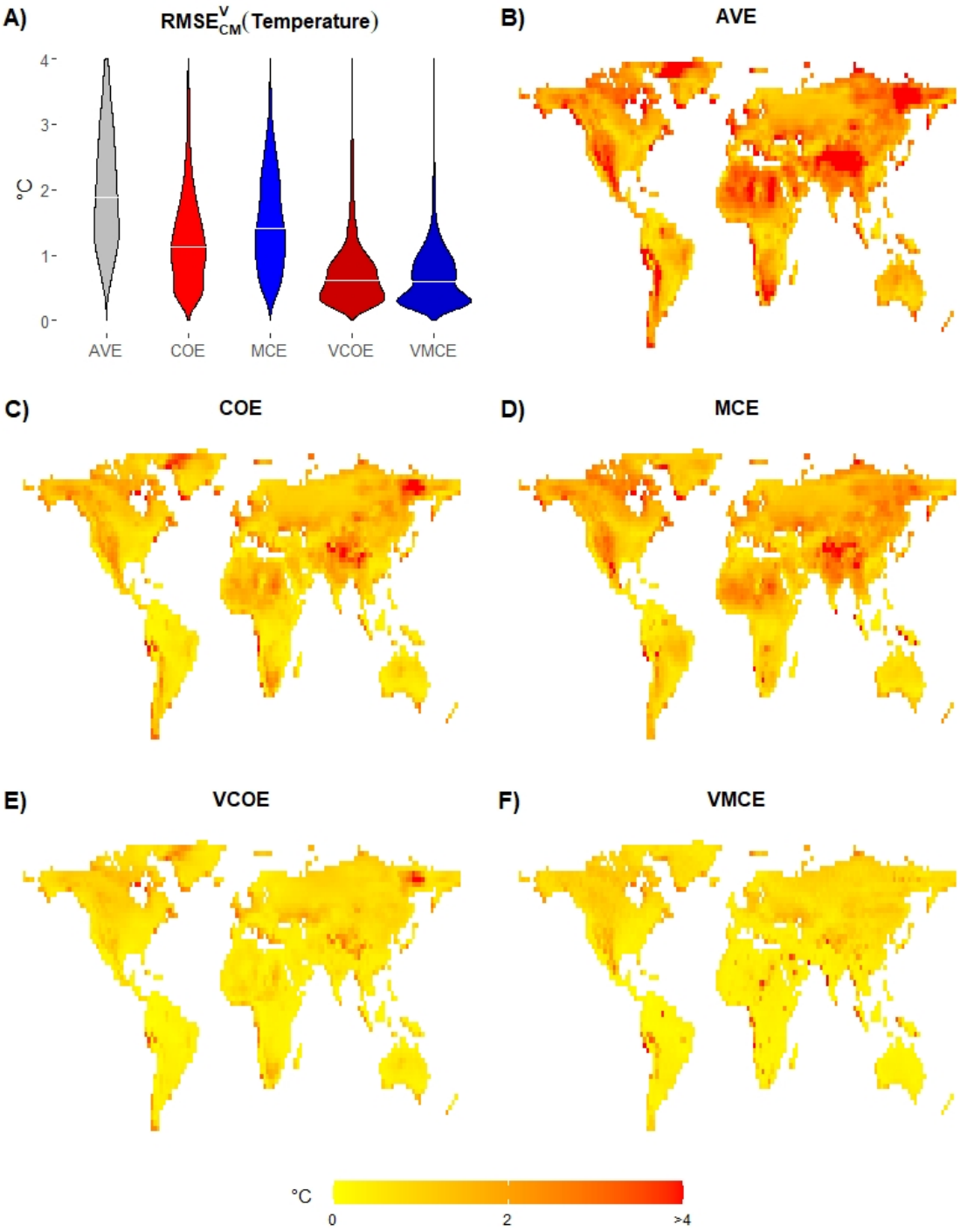

Figure 3.11: Climatological monthly RMSE ($RMSE_{CM}$) results on validation period **(years 1981-2020)** for temperature. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 4.0. **B) - F)** Geo-spatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

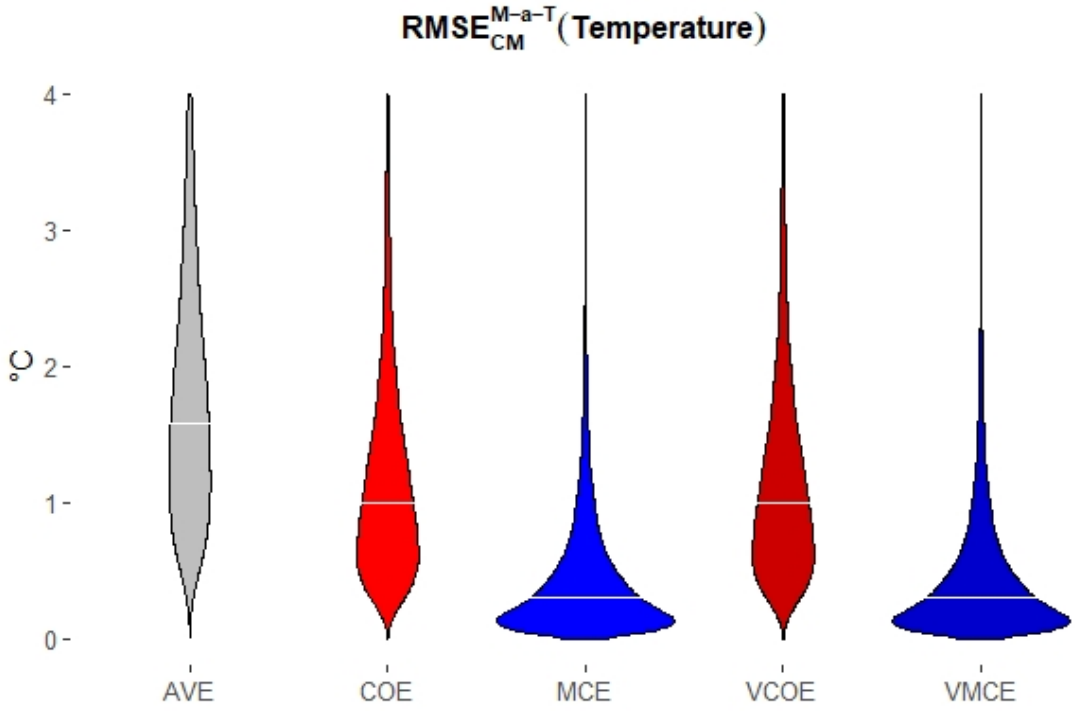

Figure 3.12: Violin plot showing model-as-truth experiment climatological monthly RMSE ($RMSE_{CM}$) results for temperature during years 2021-2100 using all land points with white lines showing median values. The Y-axis is cut at 6.0.

The VMCE method has the best $RMSE_{CM}$ performance on both training and validation periods as well as in the model-as-truth experiment. As the weights are varying between different months VCOE and VMCE are better optimised for each month individually. The same geographical location as on Figures 3.7 and 3.8 appear to be challenging for VMCE showing high contrast with the rest of the geo-spatial results. As VMCE has high $RMSE_{CM}$ performance on validation period as well as in model-as-truth experiment it indicates that VMCE is applicable for future temperature estimation in terms of minimising $RMSE_{CM}$. We analyse the B_{CM} performance of all the methods on training period (Figure 3.13), validation period (Figure 3.14) and in model-as-truth experiment (Figure 3.15) below.

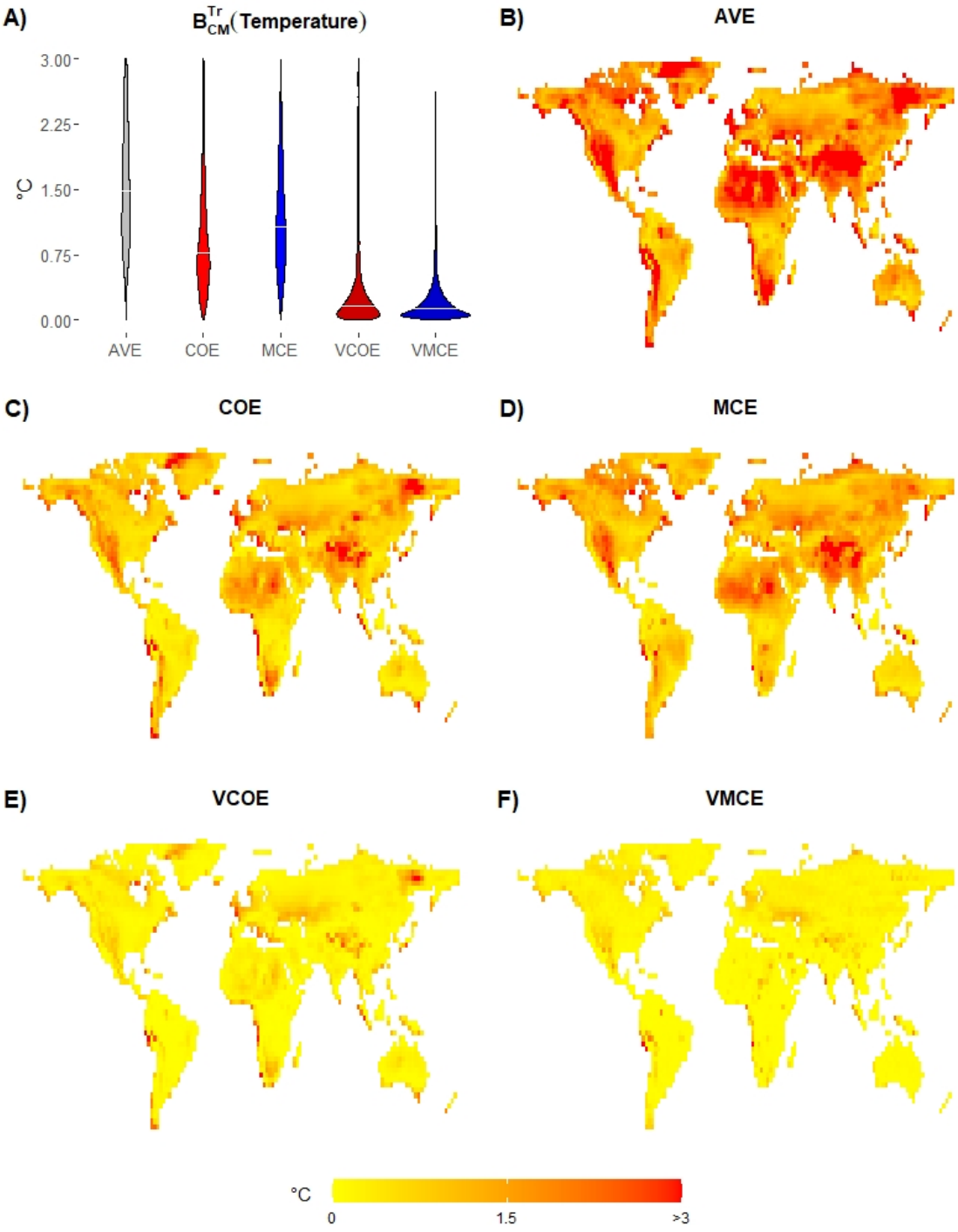

Figure 3.13: Climatological monthly bias (B_{CM}) results on training period **(years 1901-1980)** for temperature. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 3.0. **B) - F)** Geospatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

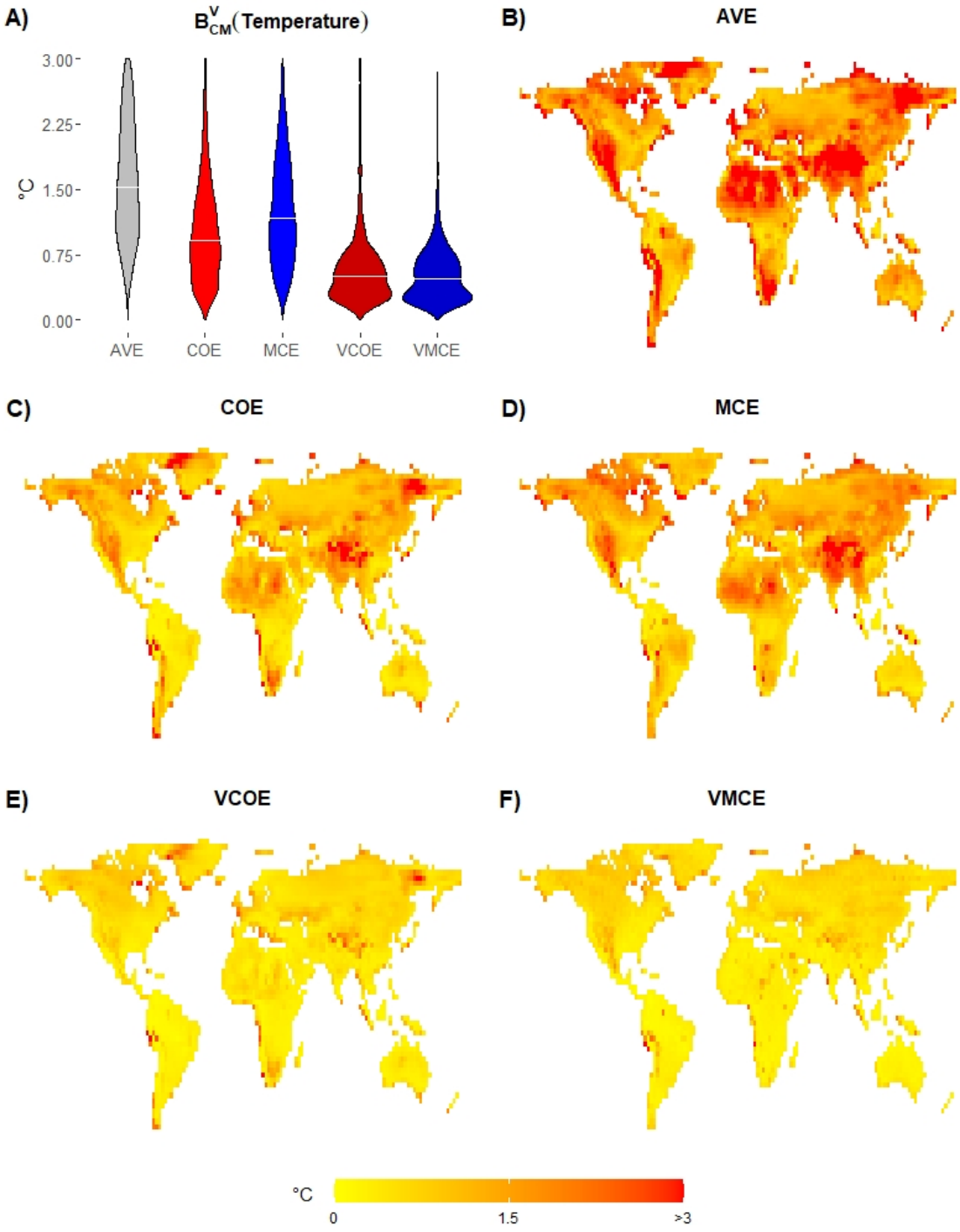

Figure 3.14: Climatological monthly bias (B_{CM}) results on validation period **(years 1981-2020)** for temperature. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 3.0. **B) - F)** Geospatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

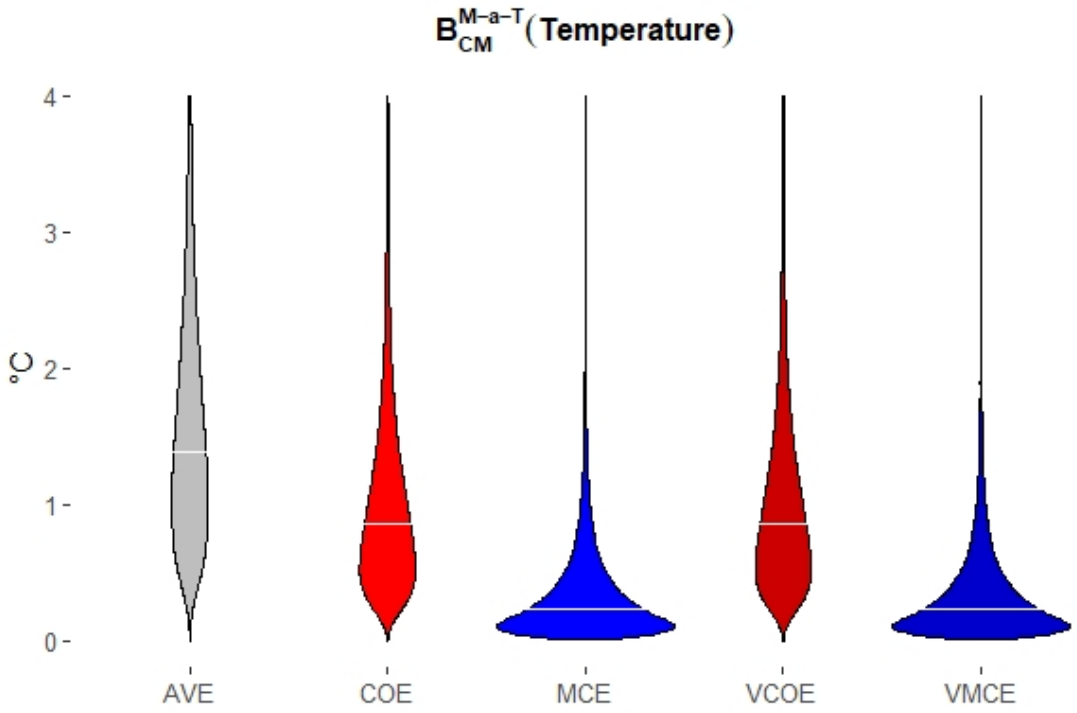

Figure 3.15: Violin plot showing model-as-truth experiment climatological monthly bias (B_{CM}) results for temperature during years 2021-2100 using all land points with white lines showing median values. The Y-axis is cut at 4.0.

The B_{CM} results are very close to the $RMSE_{CM}$ ones and confirm the findings discussed above. As VMCE has high B_{CM} performance on validation period as well as in model-as-truth experiment it indicates that VMCE is applicable for future temperature estimation in terms of minimising B_{CM}. We analyse the B_T performance of all the methods on training period (Figure 3.16), validation period (Figure 3.16) and in model-as-truth experiment (Figure 3.18) below.

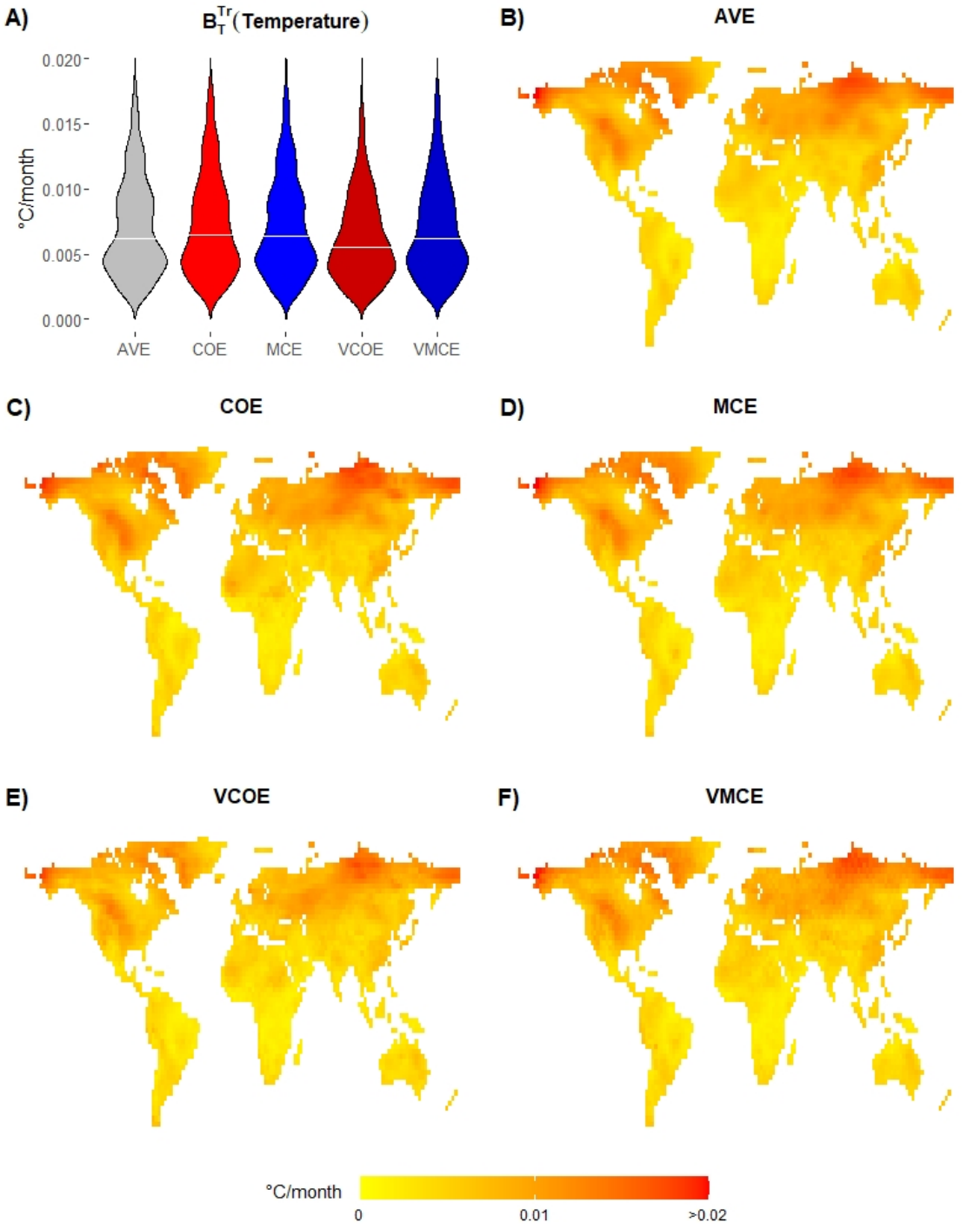

Figure 3.16: Trend bias (B_T) results on training period **(years 1901-1980)** for temperature. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 0.02. **B) - F)** Geo-spatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

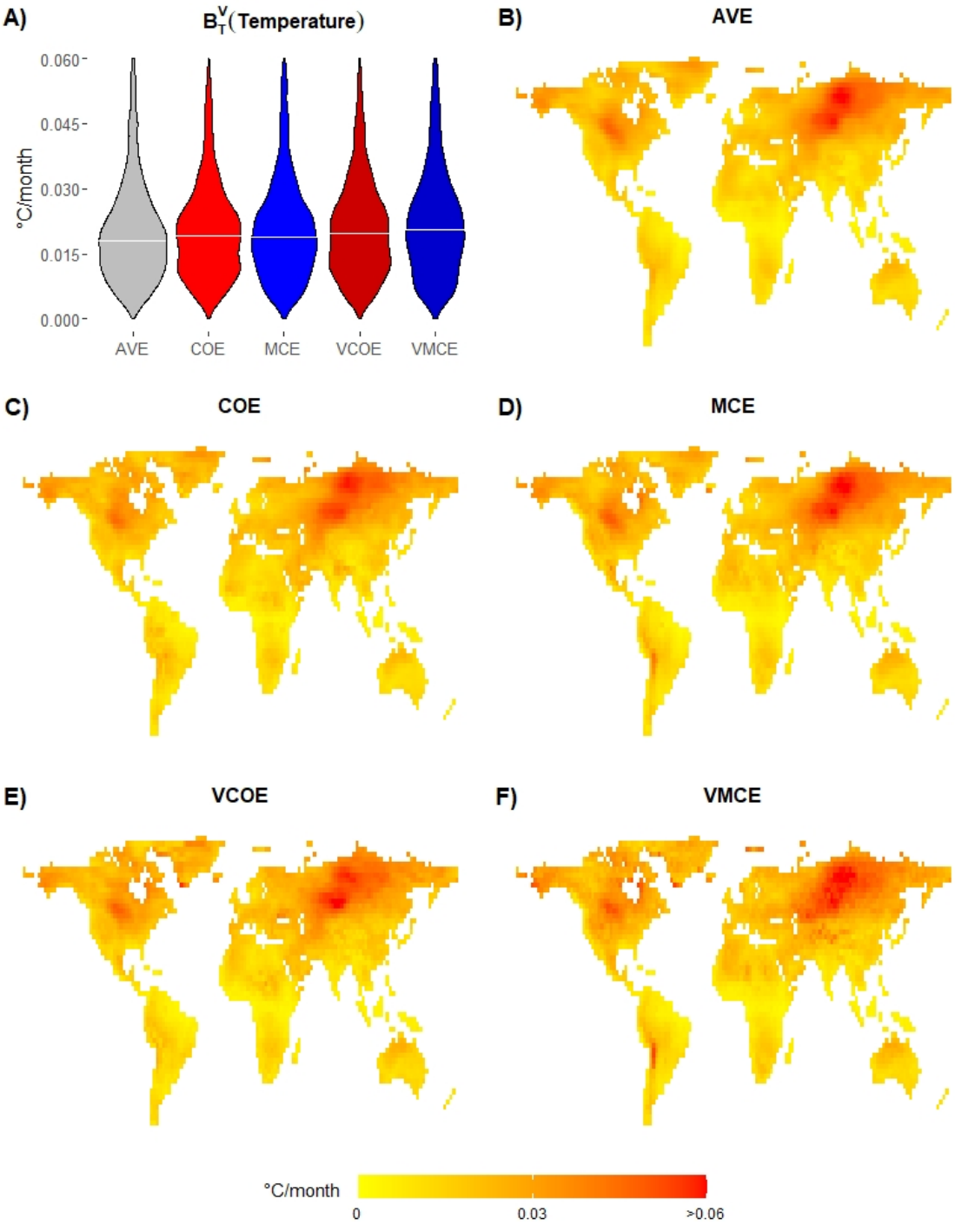

Figure 3.17: Trend bias (B_T) results on validation period **(years 1981-2020)** for temperature. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 0.06. **B) - F)** Geo-spatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

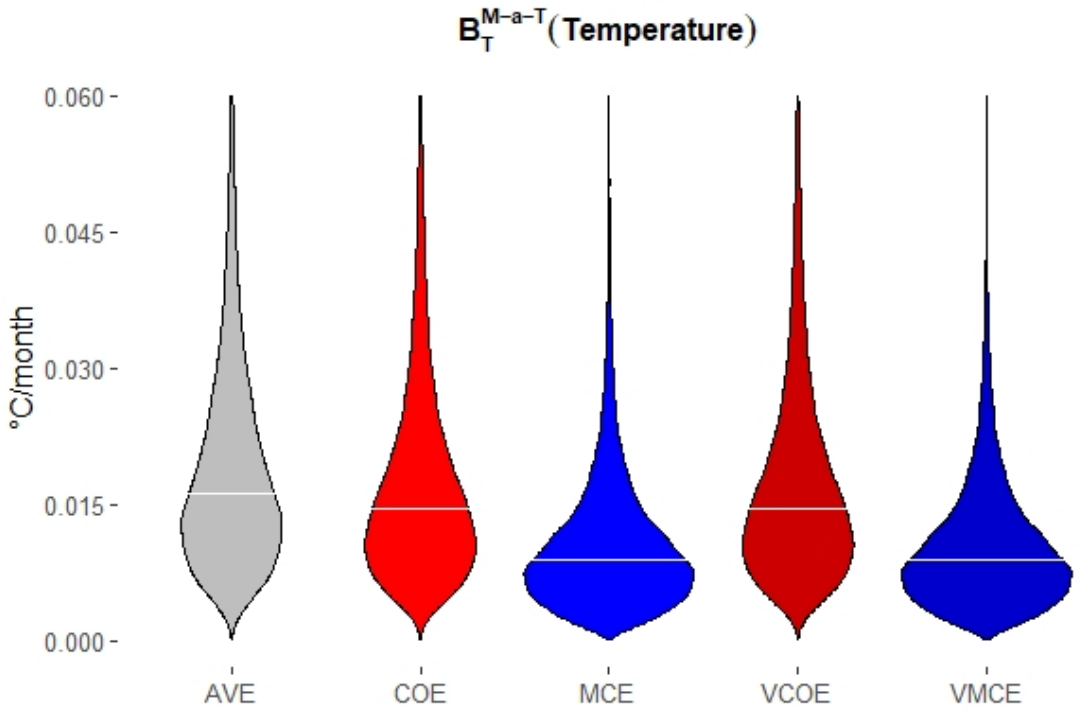

Figure 3.18: Violin plot showing model-as-truth experiment trend bias (B_T) results for temperature during years 2021-2100 using all land points with white lines showing median values. The Y-axis is cut at 0.06.

All the methods perform at similar level of B_T on training and validation data. There is no clear geo-spatial pattern where one method has advantage. This indicates that it is not possible to differentiate between methods used in this study in terms of minimising B_T on training and validation period. All the methods perform at a similar level in model-as-truth experiment with MCE and VMCE having a slight advantage. As VMCE has high B_T performance in model-as-truth experiment and the same B_T performance as the other methods on validation data it indicates that VMCE is applicable for future temperature estimation in terms of minimising B_T. We analyse the B_{IV} performance of all the methods on training period (Figure 3.19), validation period (Figure 3.20) and in model-as-truth experiment (Figure 3.21) below.

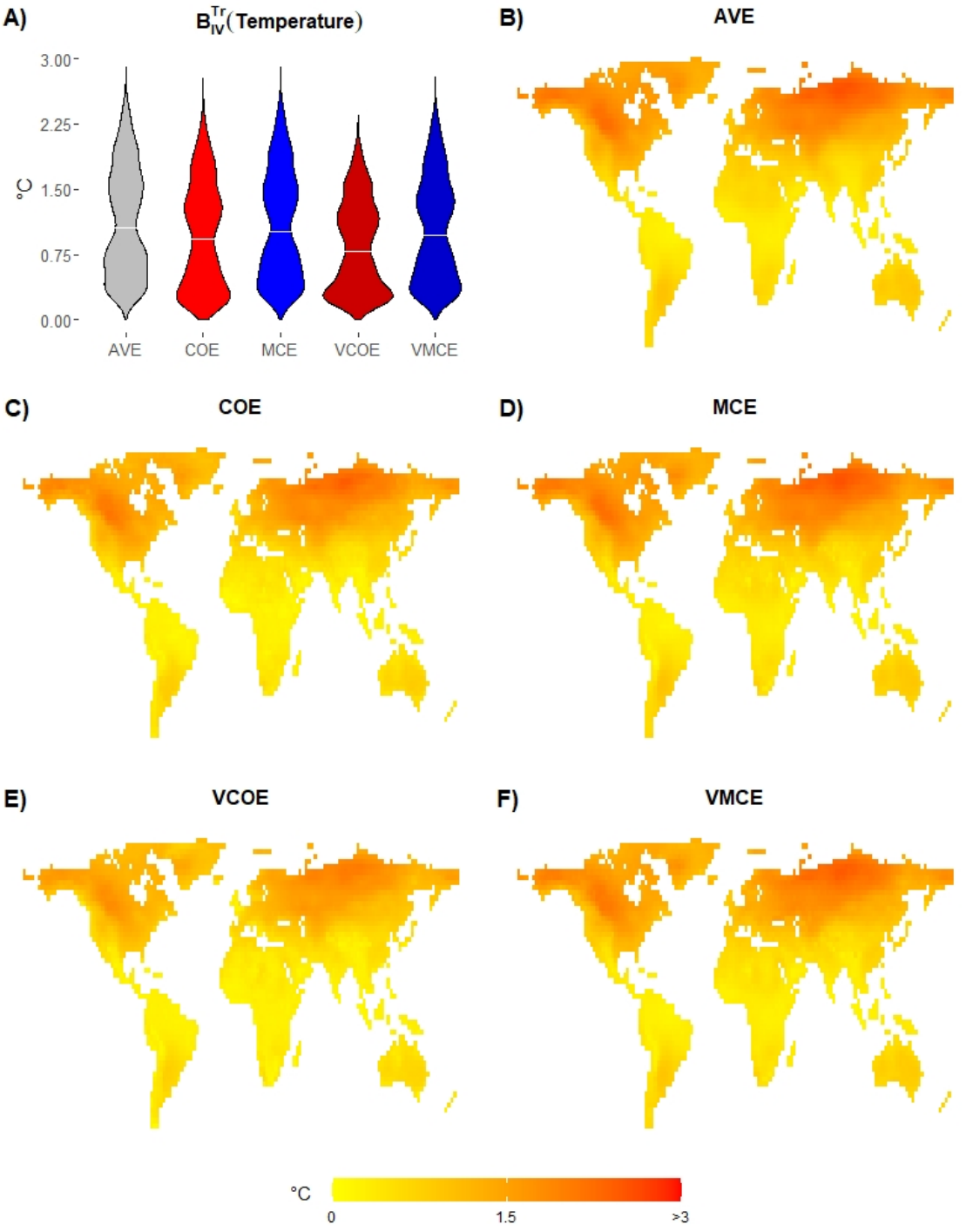

Figure 3.19: Interannual variability (B_{IV}) results on training period **(years 1901-1980)** for temperature. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 3.0. **B) - F)** Geo-spatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

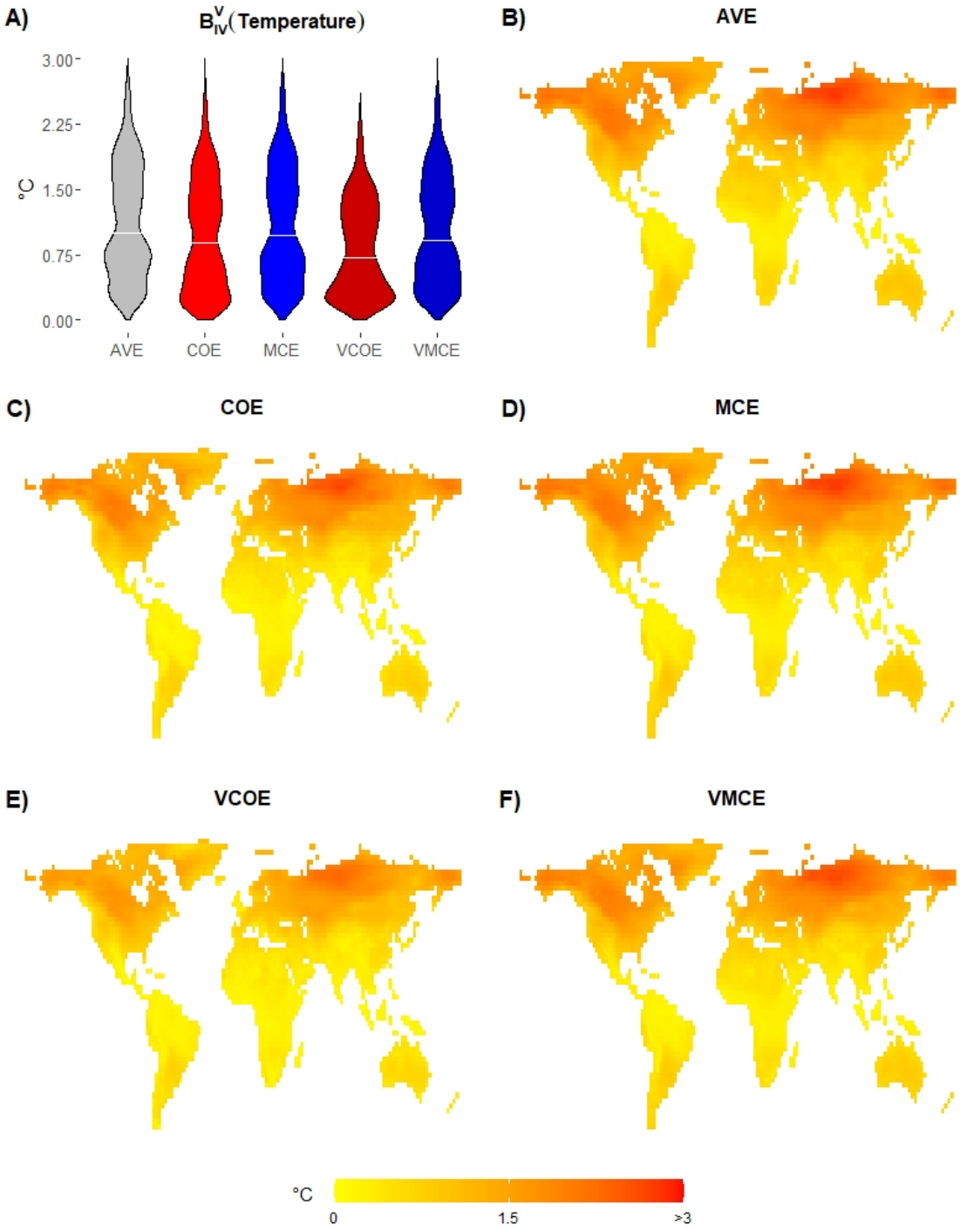

Figure 3.20: Interannual variability (B_{IV}) results on validation period **(years 1981-2020)** for temperature. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 3.0. **B) - F)** Geo-spatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

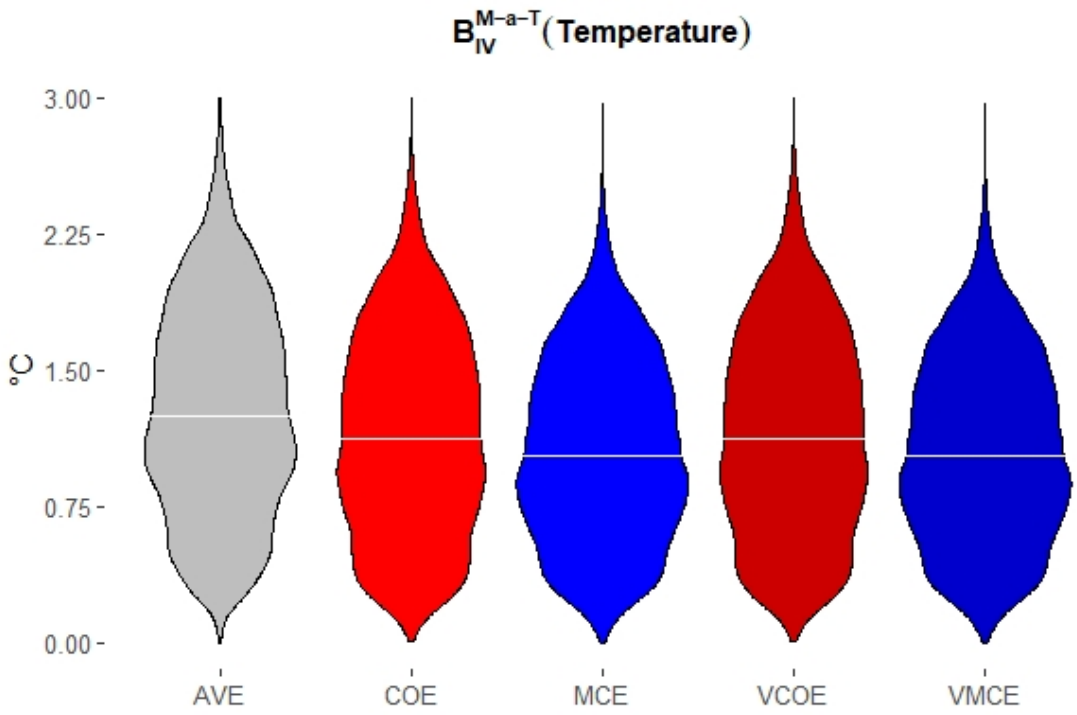

Figure 3.21: Violin plot showing model-as-truth experiment interannual variability (B_{IV}) results for temperature during years 2021-2100 using all land points with white lines showing median values. The Y-axis is cut at 3.0.

Though all the methods have similar level of B_{IV} performance, COE and VMCE have a slight advantage as their variation is closer to a single model output's variation and observations' variation. The COE method is performing best on both training and validation periods as it is excluding a number of models by assigning 0 weights as shown in Figures 2.3, 2.5 and 2.7. The VMCE method has slightly higher variation than AVE and MCE due to having more degrees of freedom from varying process described in table 3.2. All the methods perform at a similar level in model-as-truth experiment with MCE and VMCE having a slight advantage. As VMCE has lower B_{IV} performance than VCOE on validation data and the same B_{IV} performance as the other methods in model-as-truth experiment it indicates that VMCE is not the optimal method for future temperature estimation in terms of minimising B_{IV}.

3.3.2 Precipitation data

To confirm the findings in Section 3.3.1 we present the summary of the method comparison analysis for precipitation in Tables 3.6 and 3.7 below.

Ensemble	$RMSE$	$RMSE_{CM}$	B_{CM}	B_T	B_{IV}
AVE	40.44	25.84	20.56	0.10	19.93
COE	35.78	18.69	14.38	0.10	19.74
MCE	34.48	16.97	13.30	0.10	19.92
VCOE	34.15	14.69	10.18	0.10	19.52
VMCE	**28.85**	**4.15**	**2.71**	0.10	**19.04**

Table 3.6: Average precipitation results using all land points weighted according to their area sizes on training period **(years 1901-1980)**. The minimum values in each column are emphasised in bold.

As the precipitation data is different from the temperature data in terms of distribution, model correlation and other properties (non-Gaussian distribution with low correlation between ensemble models as shown in Figures 3.2 and 3.4) the training results show significantly higher advantage of the VMCE method especially in $RMSE_{CM}$ and B_{CM} metrics. This advantage is maintained on validation data as shown in Table 3.5 below.

Ensemble	$RMSE$	$RMSE_{CM}$	B_{CM}	B_T	B_{IV}
AVE	41.95	25.29	20.19	0.34	21.96
COE	37.68	18.64	14.43	0.34	21.73
MCE	36.66	17.24	13.51	0.34	21.94
VCOE	36.22	15.59	11.31	0.34	21.53
VMCE	**32.84**	**8.21**	**6.08**	0.34	**20.95**

Table 3.7: Average precipitation results using all land points weighted according to their area sizes on validation period **(years 1981-2020)**. The minimum values in each column are emphasised in bold.

The training and validation results on precipitation data are consistent with temperature data results and with findings in Chapter 2. Both MCE and VMCE methods are showing their best performance on non-negative data with non-normal distribution of model outputs and observations (as discussed in Sections 2.3.3).

The detailed results for each climatological metric are presented as maps and distributions of the results for different methods on training and validation periods together with model-as-truth experiments' results (see Figures 3.7, 3.8, 3.9 for $RMSE$, Figures 3.10, 3.11, 3.12 for $RMSE_{CM}$, Figures 3.13, 3.14, 3.15 for B_{CM}, Figures 3.16, 3.17, 3.18 for B_T and Figures 3.19, 3.20, 3.21 for B_{IV}). The violin plots describe the distribution of metric values for each method using all land points with white lines showing median values. The violin plots title describes the metric displayed on violin and spatial plots in form of $[Metric\ abbreviation]^{[Time\ period]}(Precipitation)$ with metric abbreviation being $RMSE$ for root mean squared error, $RMSE_{CM}$ for climatological monthly RMSE, B_{CM} for climatological monthly bias, B_T for trend bias and B_{IV} for interannual variability; time period being Tr for training period (years 1901-1980), V for validation period (years 1981-2020) and $M-a-T$ for model-as-truth experiment period (years 2021-2100). The spatial plots describe the spatial distribution of metric values for each method with according method names in title. All the plots exclude small portions of data which constitute long tails of the data distributions. The maximum included values are specified in violin plots' axis and in the captions. Darker colours indicate geographical locations with higher metric values

We analyse the $RMSE$ performance of all the methods on training period (Figure 3.22), validation period (Figure 3.23) and in model-as-truth experiment (Figure 3.24) below.

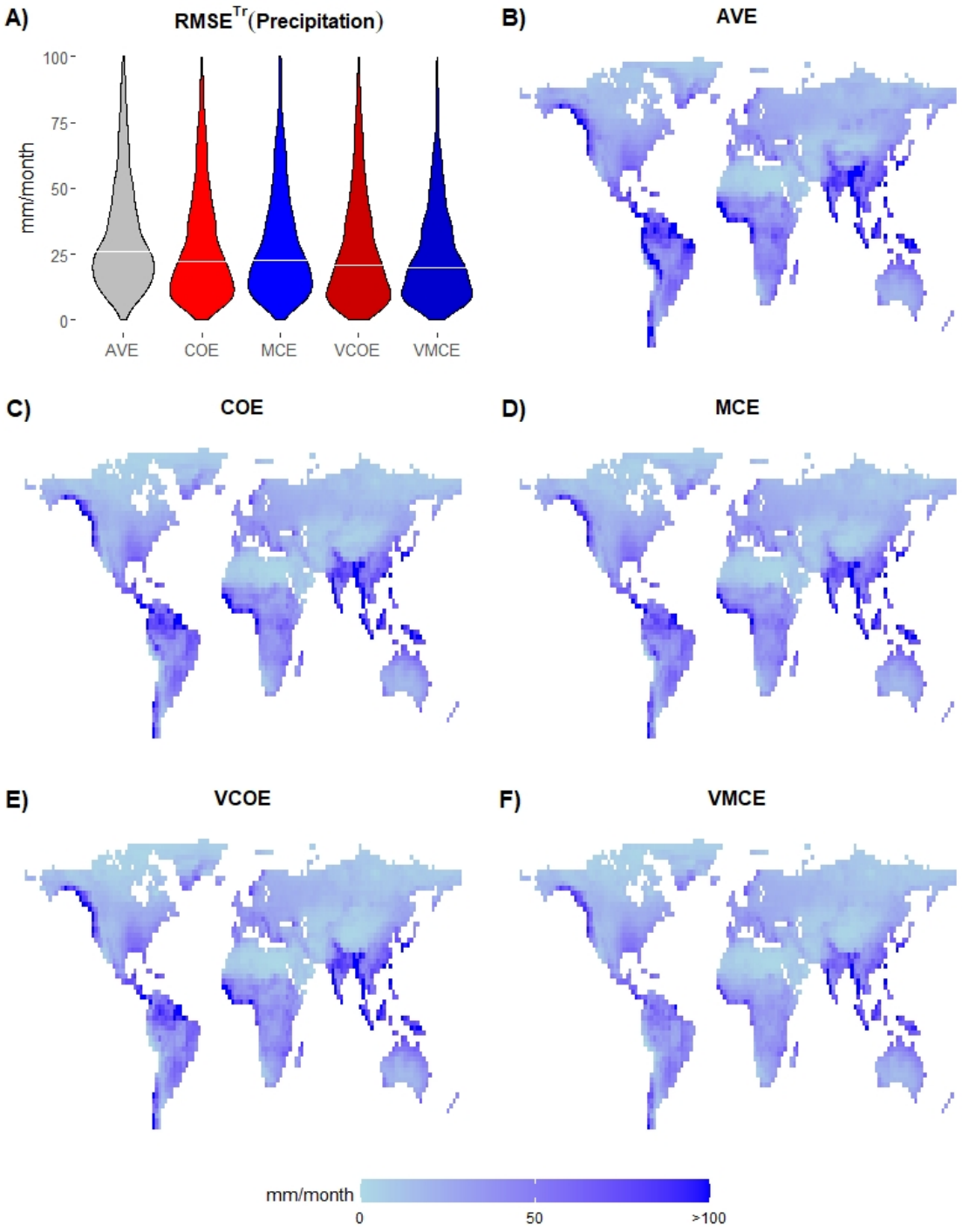

Figure 3.22: Root mean squared error ($RMSE$) results on training period **(years 1901-1980)** for precipitation. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 100. **B) - F)** Geospatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

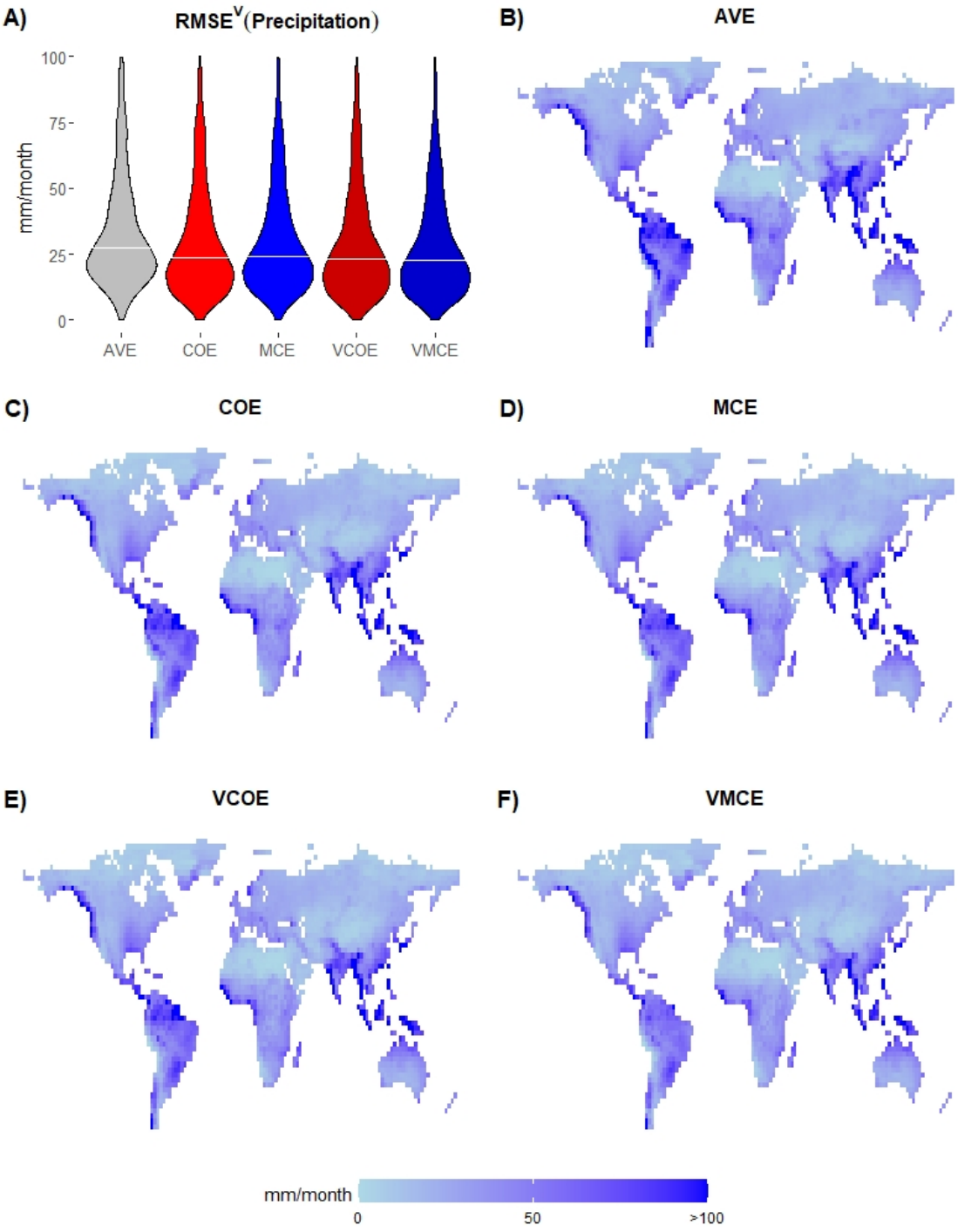

Figure 3.23: Root mean squared error ($RMSE$) results on validation period **(years 1981-2020)** for precipitation. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 100. **B) - F)** Geospatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

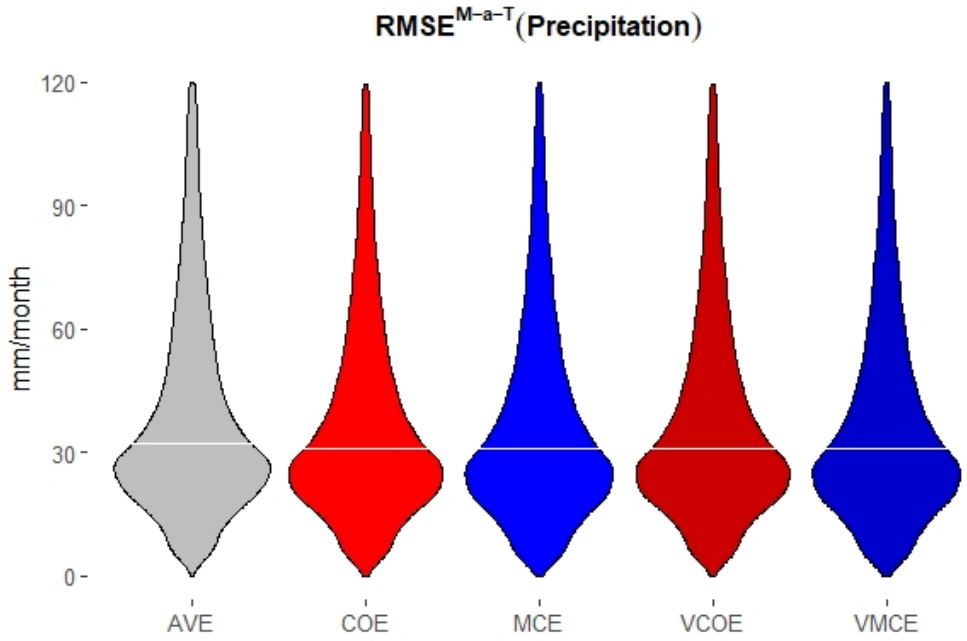

Figure 3.24: Violin plot showing model-as-truth experiment root mean squared error ($RMSE$) results for precipitation during years 2021-2100 using all land points with white lines showing median values. The Y-axis is cut at 120.

The precipitation data results confirm the findings from the temperature data results with the VMCE method having slightly higher $RMSE$ performance than all other methods on both training and validation periods. This indicates that on non-Gaussian distributed data with weakly correlated climate models both the MCE and the VMCE methods are able to find more optimal weights compared to the COE and the VCOE method respectively even on training data with the default sets of parameters. The areas of highest $RMSE$ values are similar for all the methods with the VMCE method being able to find an optimal solution for some of the grid cells which are challenging for other methods. All the methods perform at a similar level in model-as-truth experiment. As VMCE has high $RMSE$ performance on validation data and the same $RMSE$ performance as the other methods in model-as-truth experiment it indicates that VMCE is applicable for future precipitation estimation in terms of minimising $RMSE$. We analyse the $RMSE_{CM}$ performance of all the methods on training period (Figure 3.25), validation period (Figure 3.26) and in model-as-truth experiment (Figure 3.27) below.

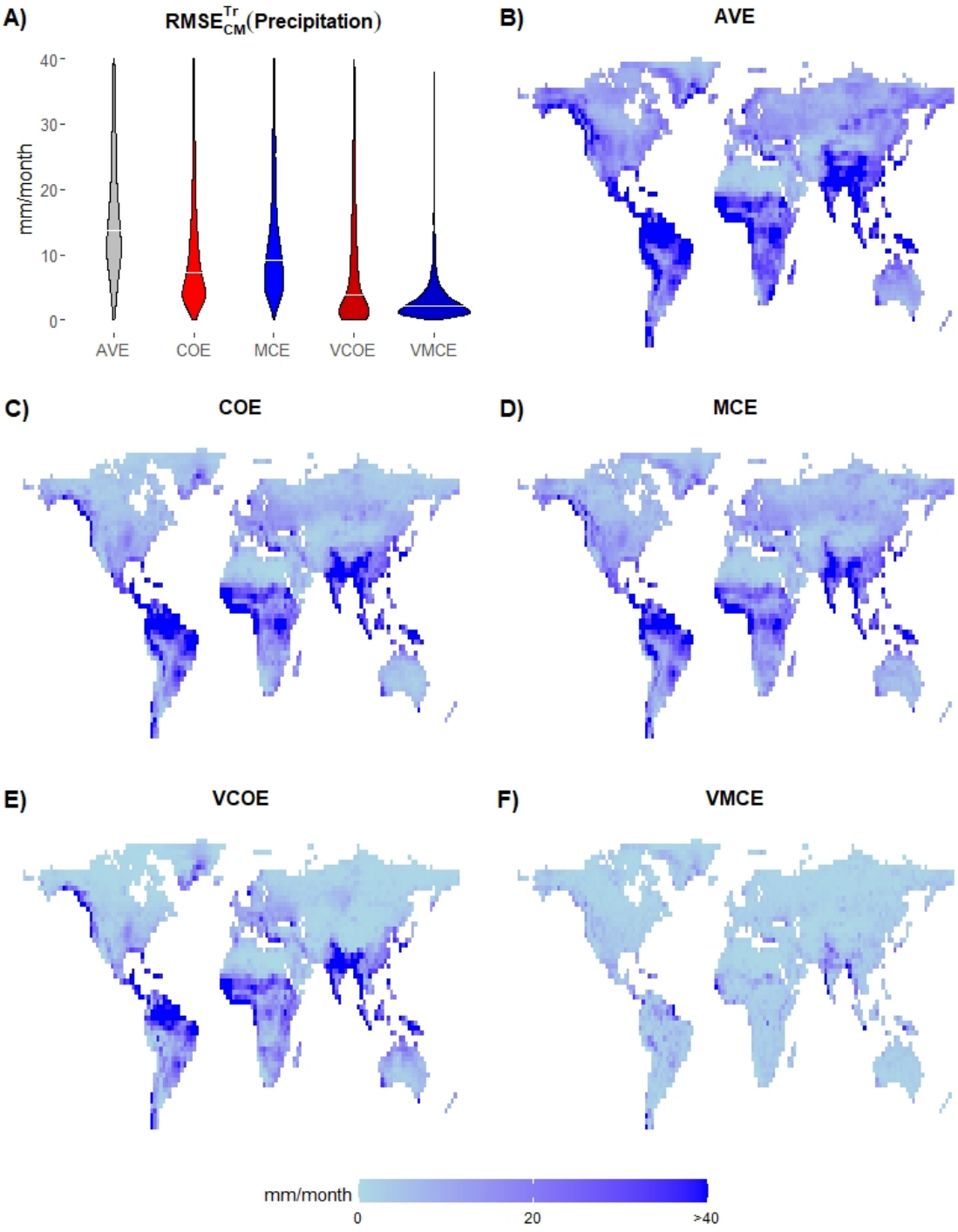

Figure 3.25: Climatological monthly RMSE ($RMSE_{CM}$) results on training period **(years 1901-1980)** for precipitation. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 40. **B) - F)** Geo-spatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

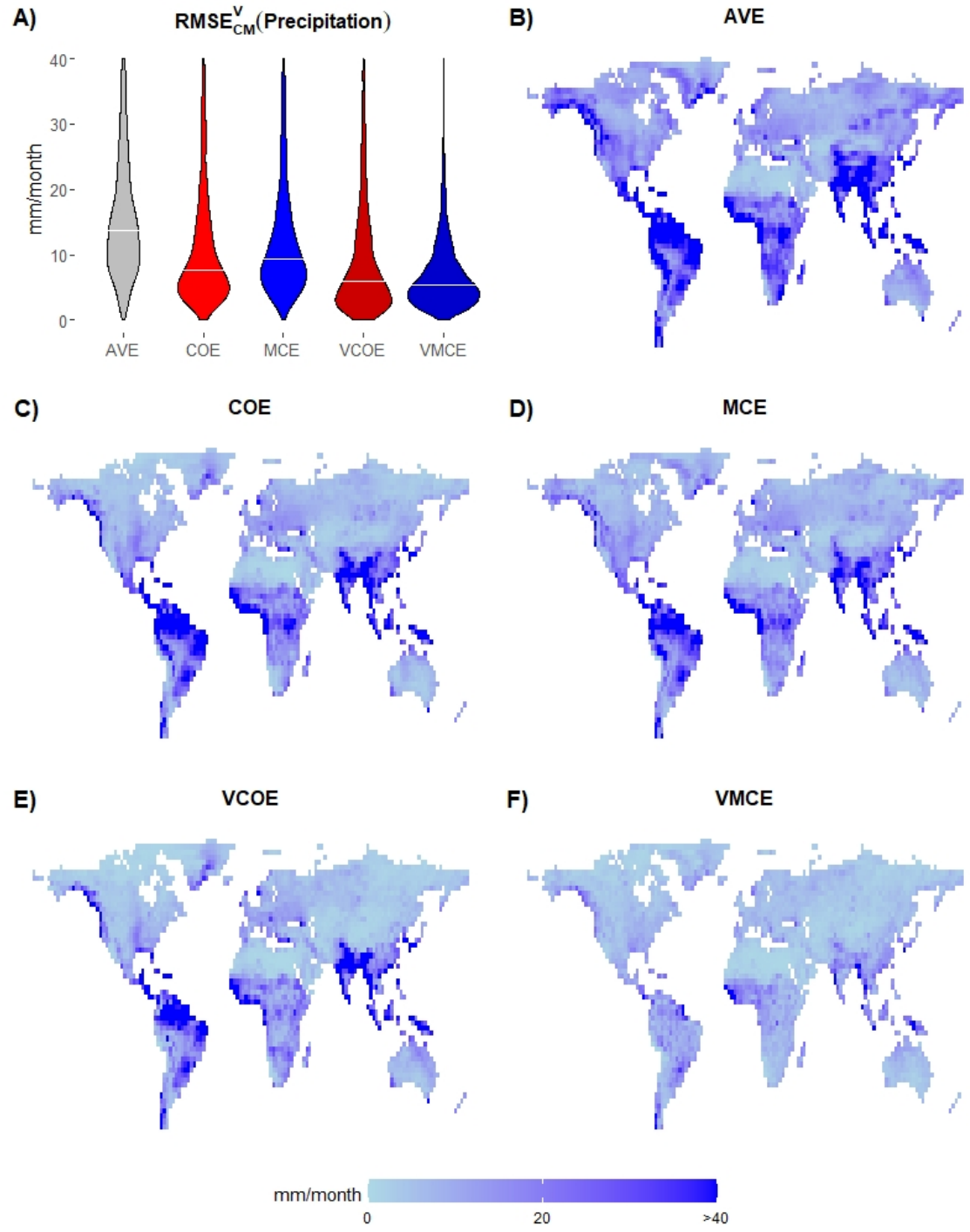

Figure 3.26: Climatological monthly RMSE ($RMSE_{CM}$) results on validation period **(years 1981-2020)** for precipitation. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 40. **B) - F)** Geo-spatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

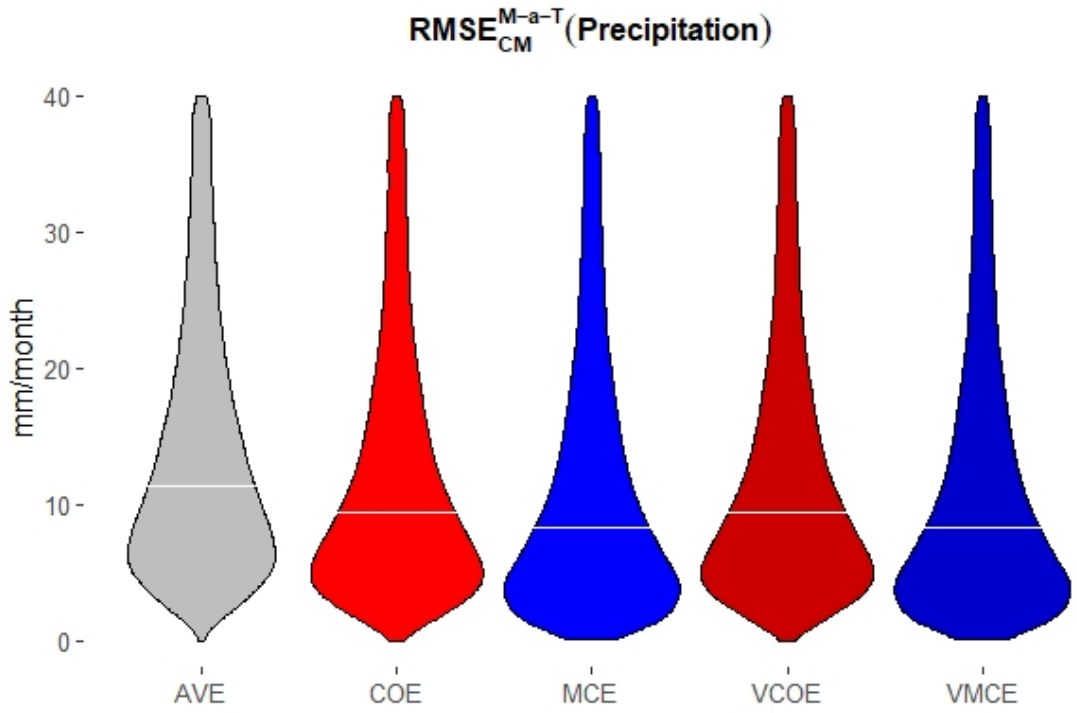

Figure 3.27: Violin plot showing model-as-truth experiment climatological monthly RMSE ($RMSE_{CM}$) results for precipitation during years 2021-2100 using all land points with white lines showing median values. The Y-axis is cut at 40.

The $RMSE_{CM}$ results confirm the previous findings. The significant difference between the VMCE and any other method indicate that the MCE method has a higher benefit from the introduction of varying weights scheme. The areas of highest $RMCE$ values are similar for all the methods except for the VMCE method which performs almost uniformly well across the whole grid. All the methods perform at a similar level in model-as-truth experiment with MCE and VMCE having a slight advantage. As VMCE has high $RMSE_{CM}$ performance on validation period as well as in model-as-truth experiment it indicates that VMCE is applicable for future precipitation estimation in terms of minimising $RMSE_{CM}$. We analyse the B_{CM} performance of all the methods on training period (Figure 3.28), validation period (Figure 3.29) and in model-as-truth experiment (Figure 3.30) below.

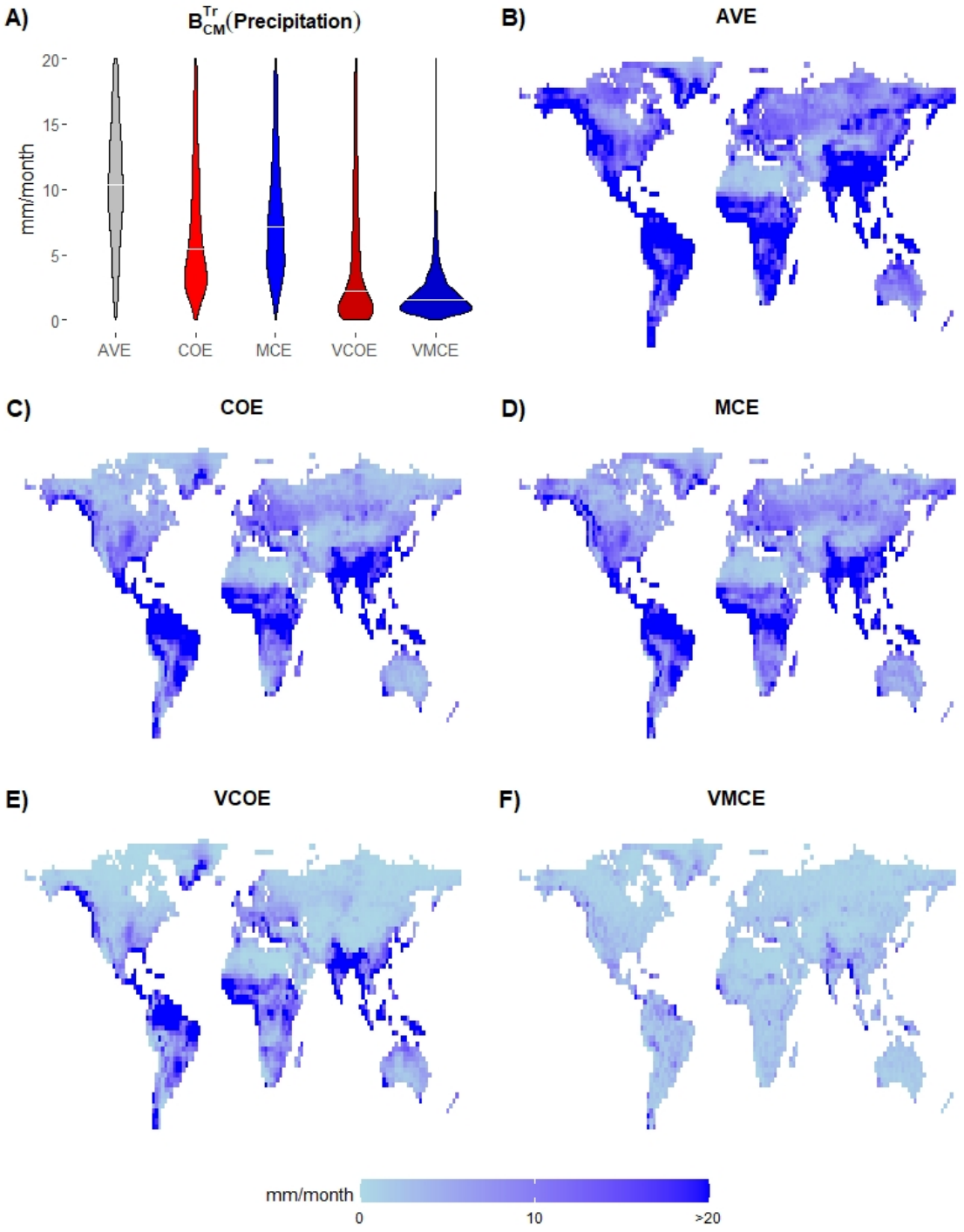

Figure 3.28: Climatological monthly bias (B_{CM}) results on training period **(years 1901-1980)** for precipitation. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 20. **B) - F)** Geospatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

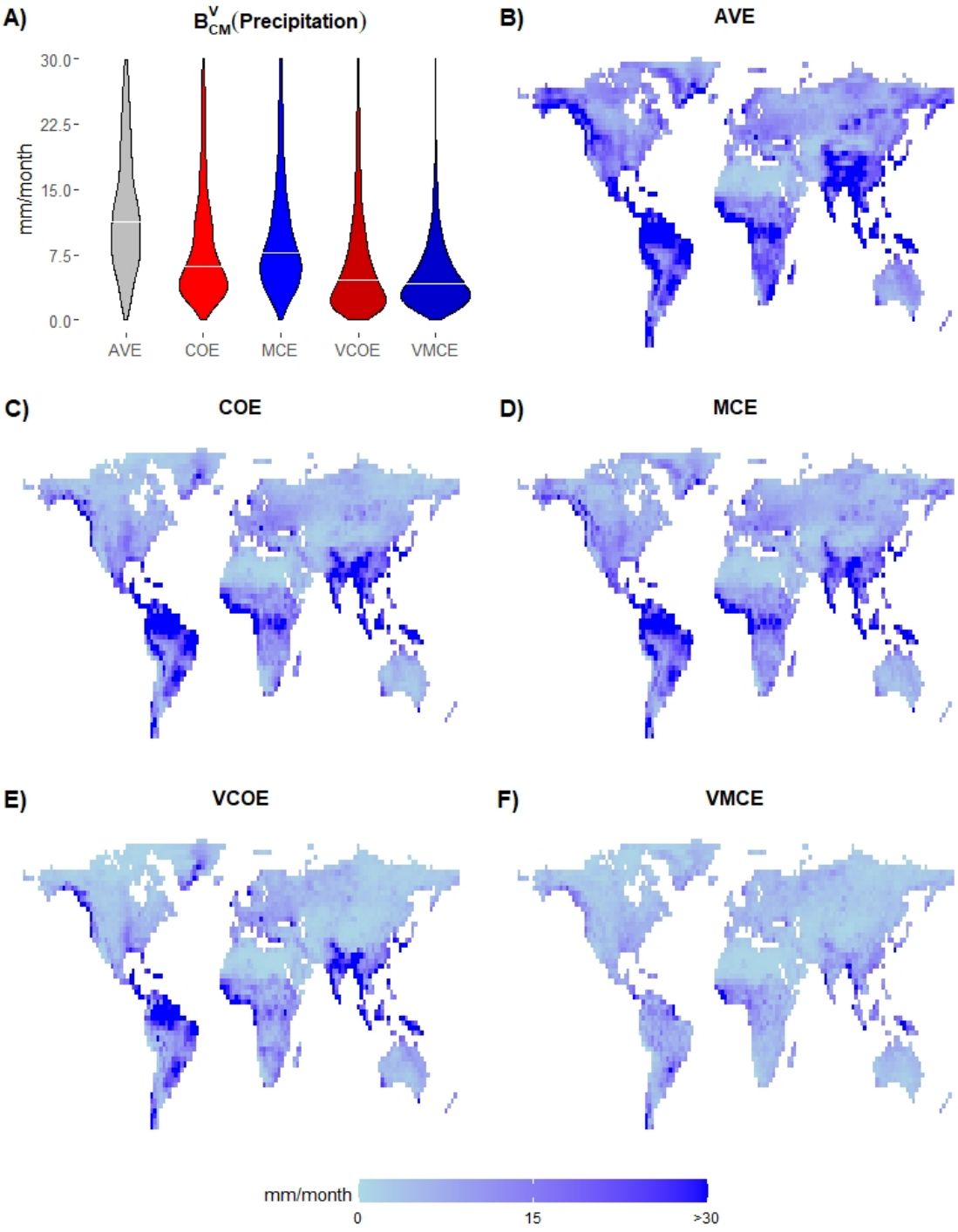

Figure 3.29: Climatological monthly bias (B_{CM}) results on validation period **(years 1981-2020)** for precipitation. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 30. **B) - F)** Geo-spatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

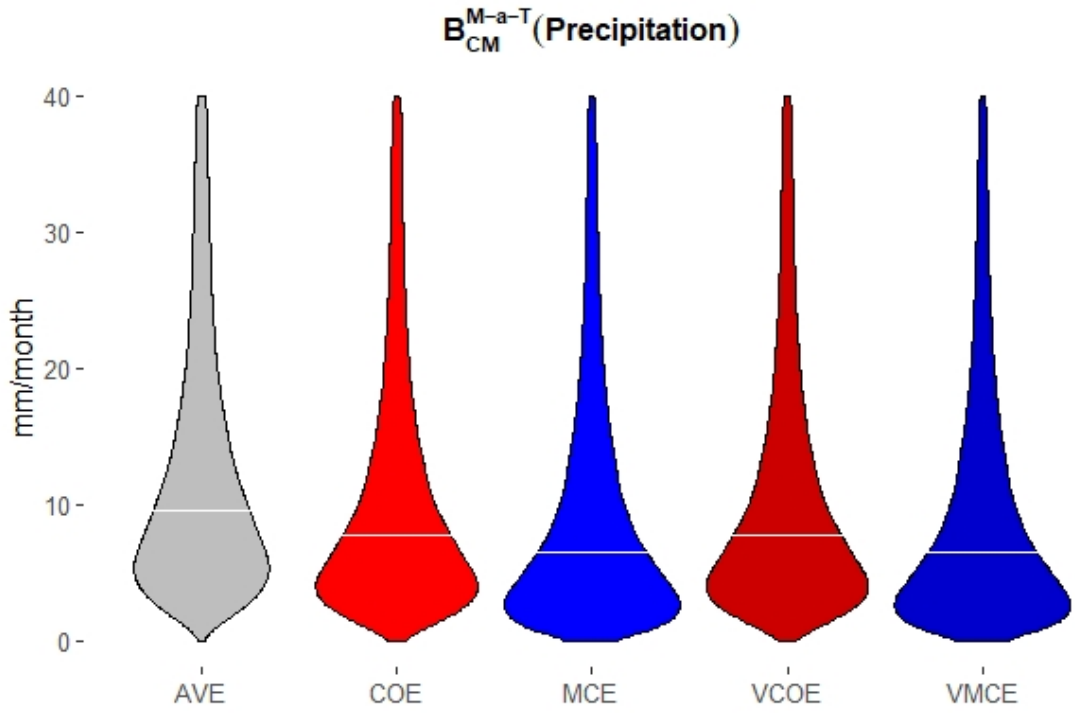

Figure 3.30: Violin plot showing model-as-truth experiment climatological monthly bias (B_{CM}) results for precipitation during years 2021-2100 using all land points with white lines showing median values. The Y-axis is cut at 40.

The B_{CM} results confirm the previous findings. As VMCE has high B_{CM} performance on validation period as well as in model-as-truth experiment it indicates that VMCE is applicable for future precipitation estimation in terms of minimising B_{CM}. We analyse the B_T performance of all the methods on training period (Figure 3.31), validation period (Figure 3.32) and in model-as-truth experiment (Figure 3.33) below.

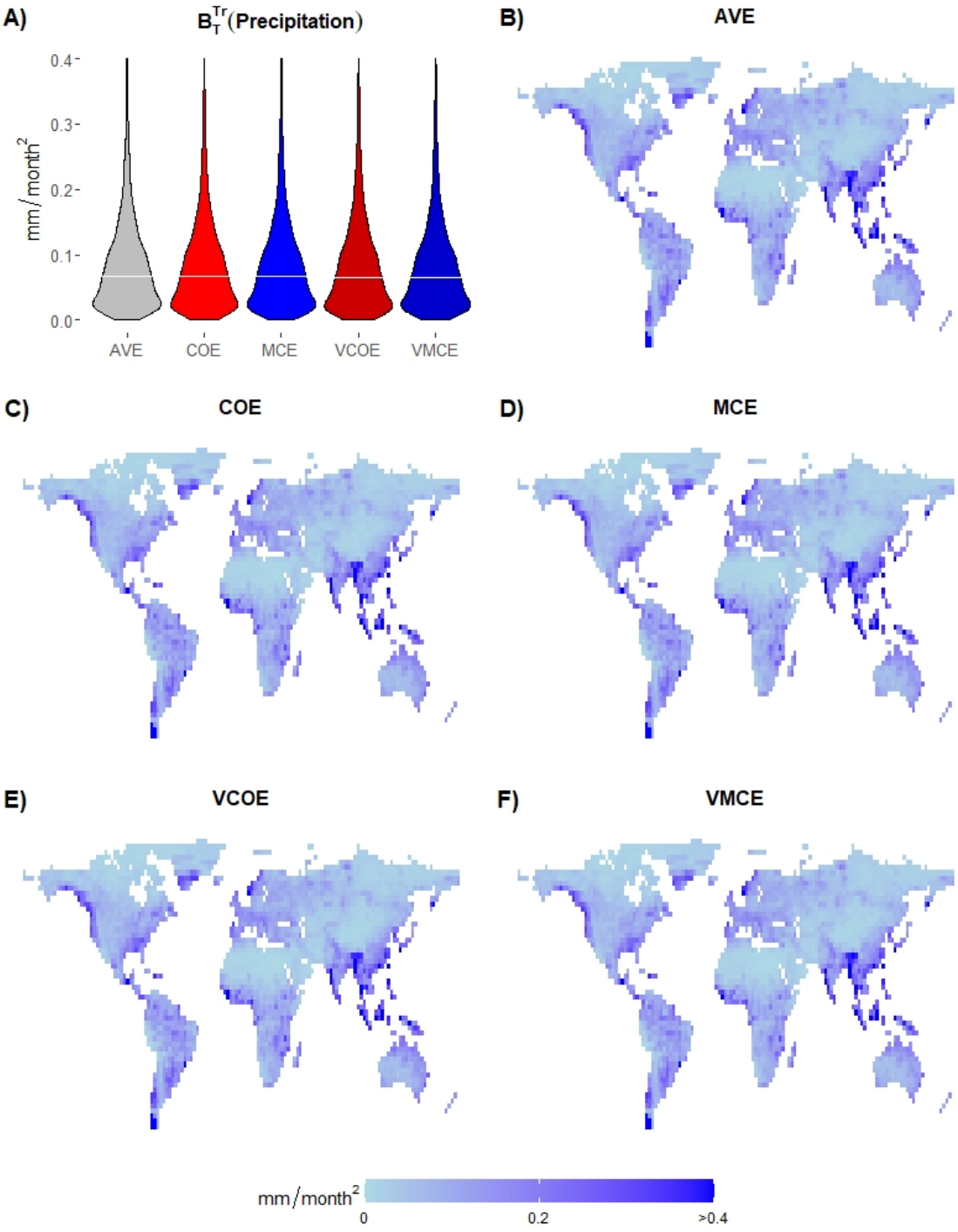

Figure 3.31: Trend bias (B_T) results on training period **(years 1901-1980)** for precipitation. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 0.4. **B) - F)** Geo-spatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

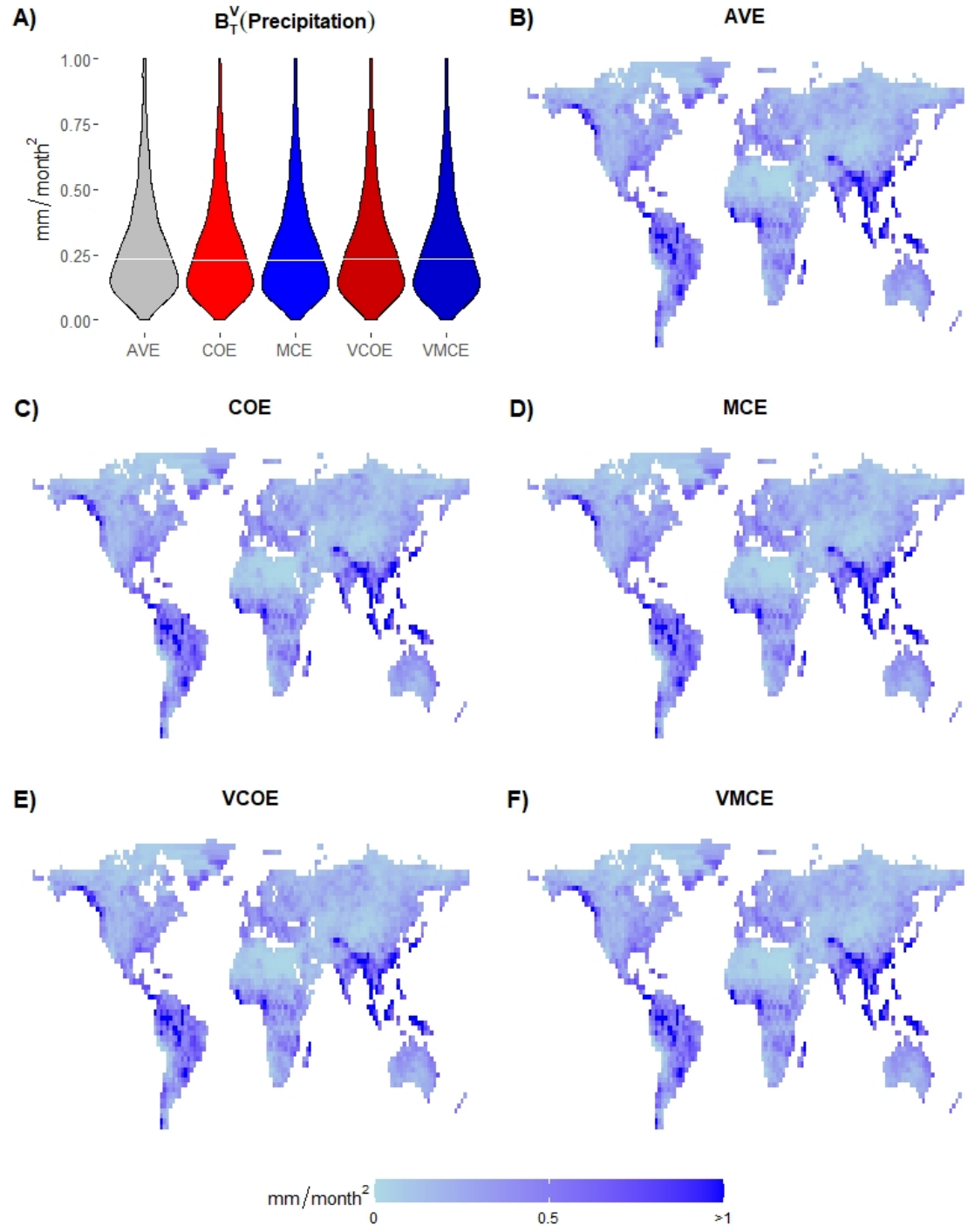

Figure 3.32: Trend bias (B_T) results on validation period **(years 1981-2020)** for precipitation. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 1.0. **B) - F)** Geo-spatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

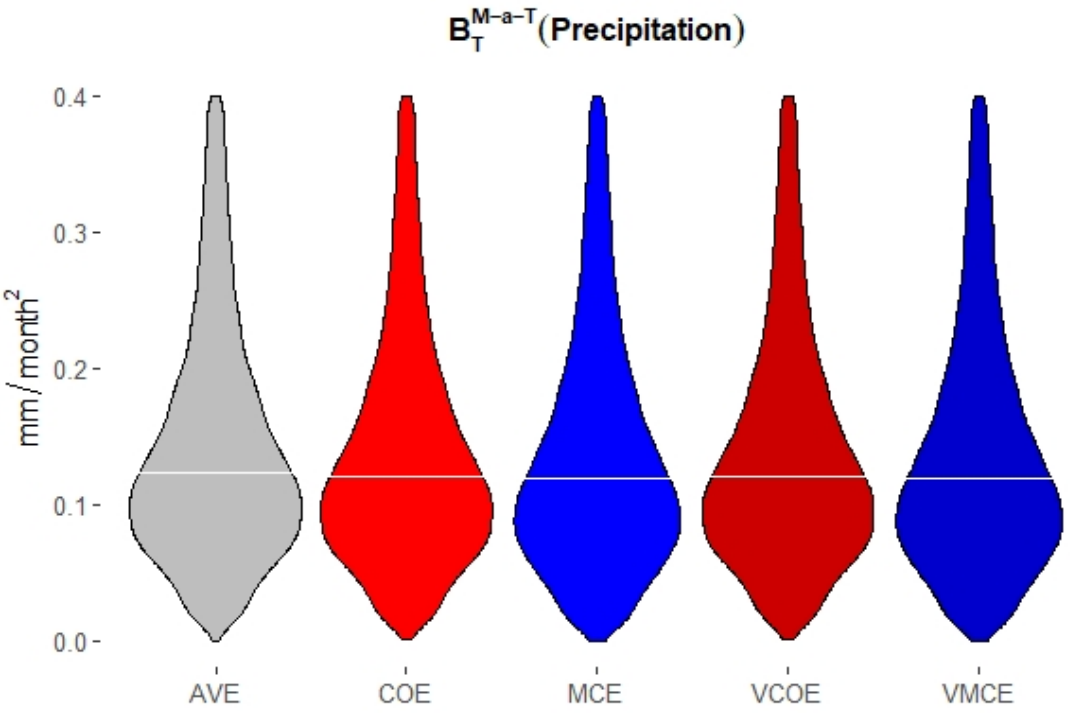

Figure 3.33: Violin plot showing model-as-truth experiment trend bias (B_T) results for precipitation during years 2021-2100 using all land points with white lines showing median values. The Y-axis is cut at 0.4.

As previously shown for the temperature data all the methods perform at similar level of B_T on training and validation data as well as in model-as-truth experiments. As VMCE has the same B_T performance on validation period and in model-as-truth experiment as other methods in this study it indicates that VMCE is as applicable for future precipitation estimation in terms of minimising B_T as other methods in this study. We analyse the B_{IV} performance of all the methods on training period (Figure 3.35), validation period (Figure 3.35) and in model-as-truth experiment (Figure 3.36) below.

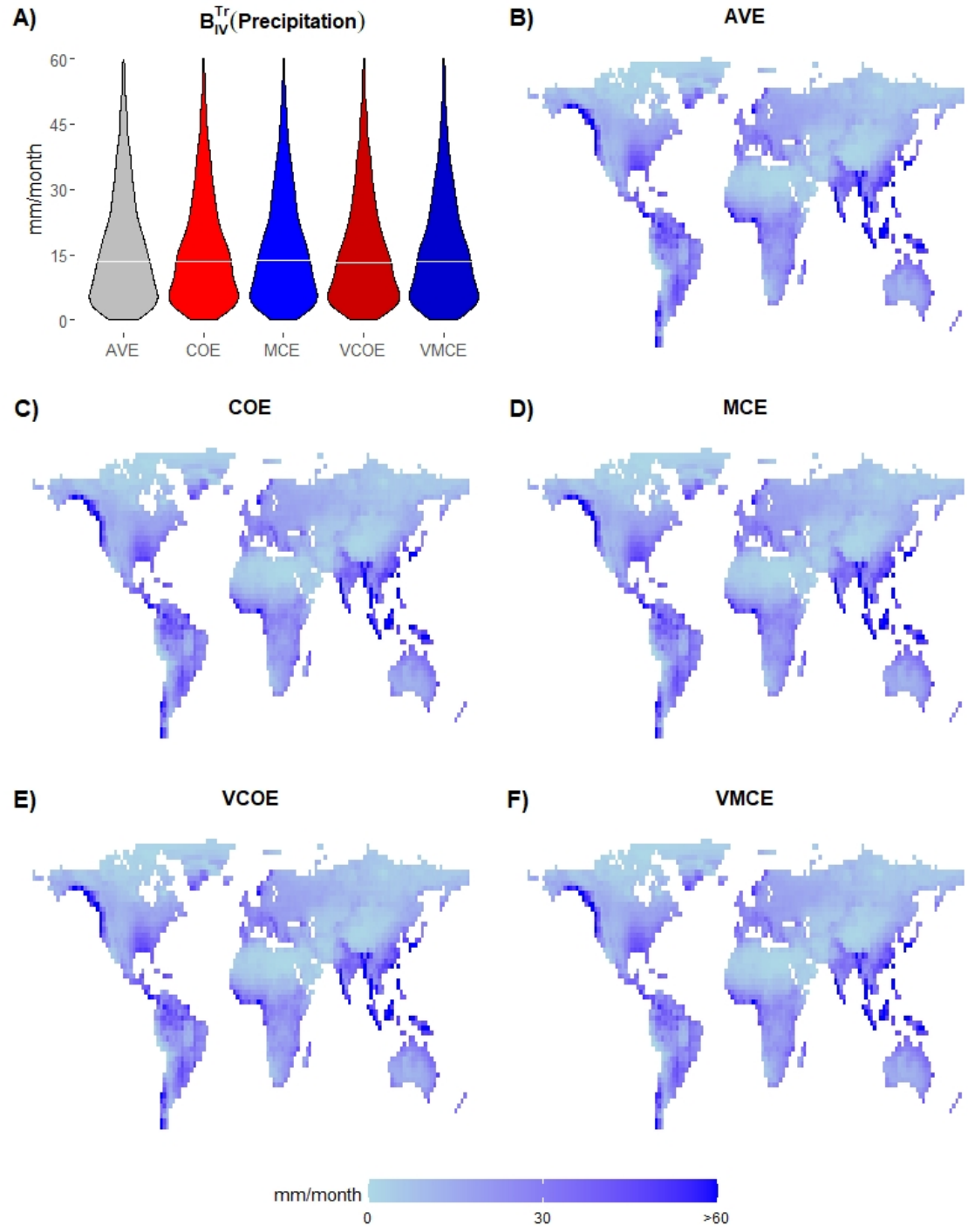

Figure 3.34: Interannual variability (B_{IV}) results on training period **(years 1901-1980)** for precipitation. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 60. **B) - F)** Geo-spatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

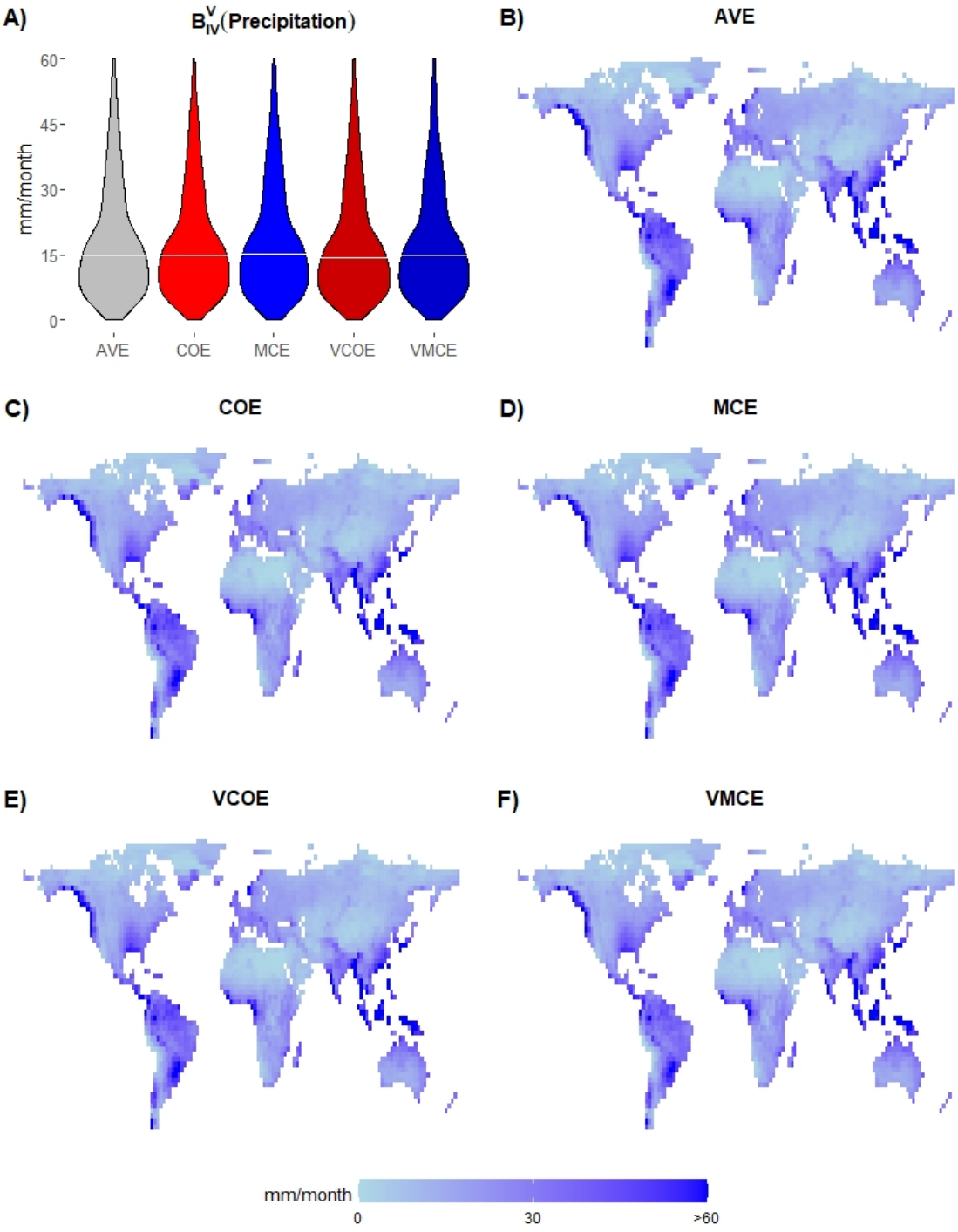

Figure 3.35: Interannual variability (B_{IV}) results on validation period **(years 1981-2020)** for precipitation. **A)** Violin plots showing distribution of the results using all land points with white lines showing median values. The Y-axis is cut at 60. **B) - F)** Geo-spatial distribution of AVE, COE, MCE, VCOE, and VMCE method results respectively.

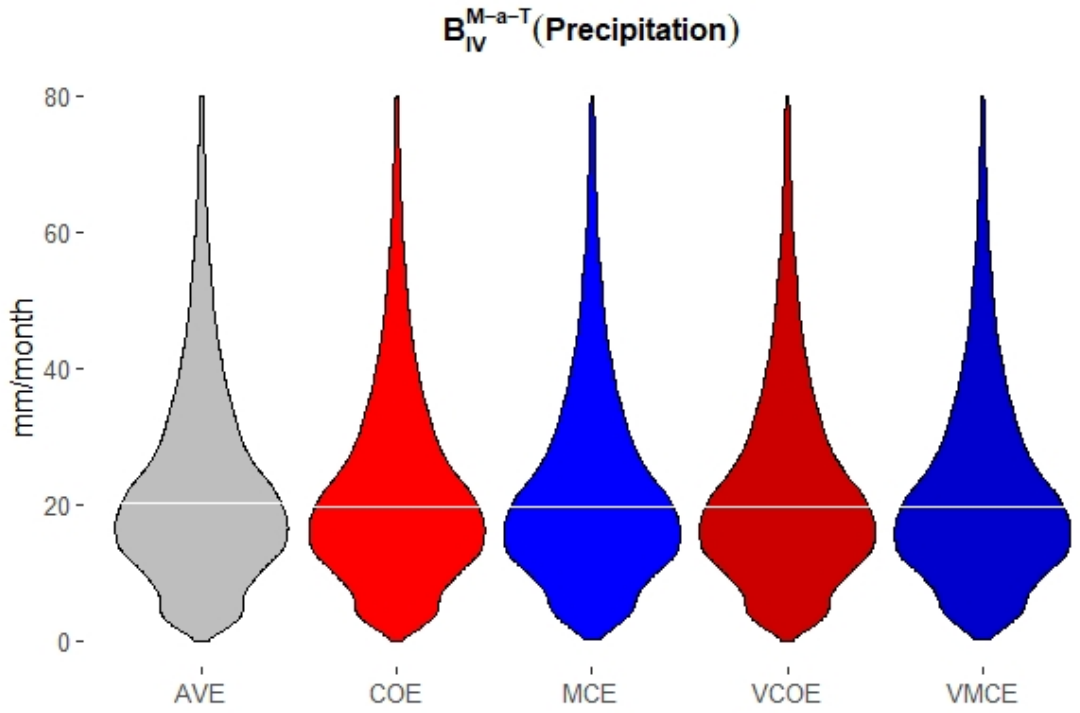

Figure 3.36: Violin plot showing model-as-truth experiment interannual variability (B_{IV}) results for precipitation during years 2021-2100 using all land points with white lines showing median values. The Y-axis is cut at 80.

All the methods have similar level of B_{IV} on training and validation periods as well as in model-as-truth experiments. As VMCE has the same B_{IV} performance on validation period and in model-as-truth experiment as other methods it indicates that VMCE is as applicable for future precipitation estimation in terms of minimising B_{IV} as other methods in this study.

3.3.3 Varying weights

To further understand the effect of introducing varying weights we analyse the VMCE method weights by looking at their distribution during year seasons in different climate zones. First we select one example grid cell in each climate zone and then we compare three sets of weights (one set of weights per month) in each year season. The results of

this analysis are presented in Figures 3.37, 3.38, 3.39 and 3.40 for temperature, and in Figures 3.41, 3.42, 3.43 and 3.44 for precipitation below.

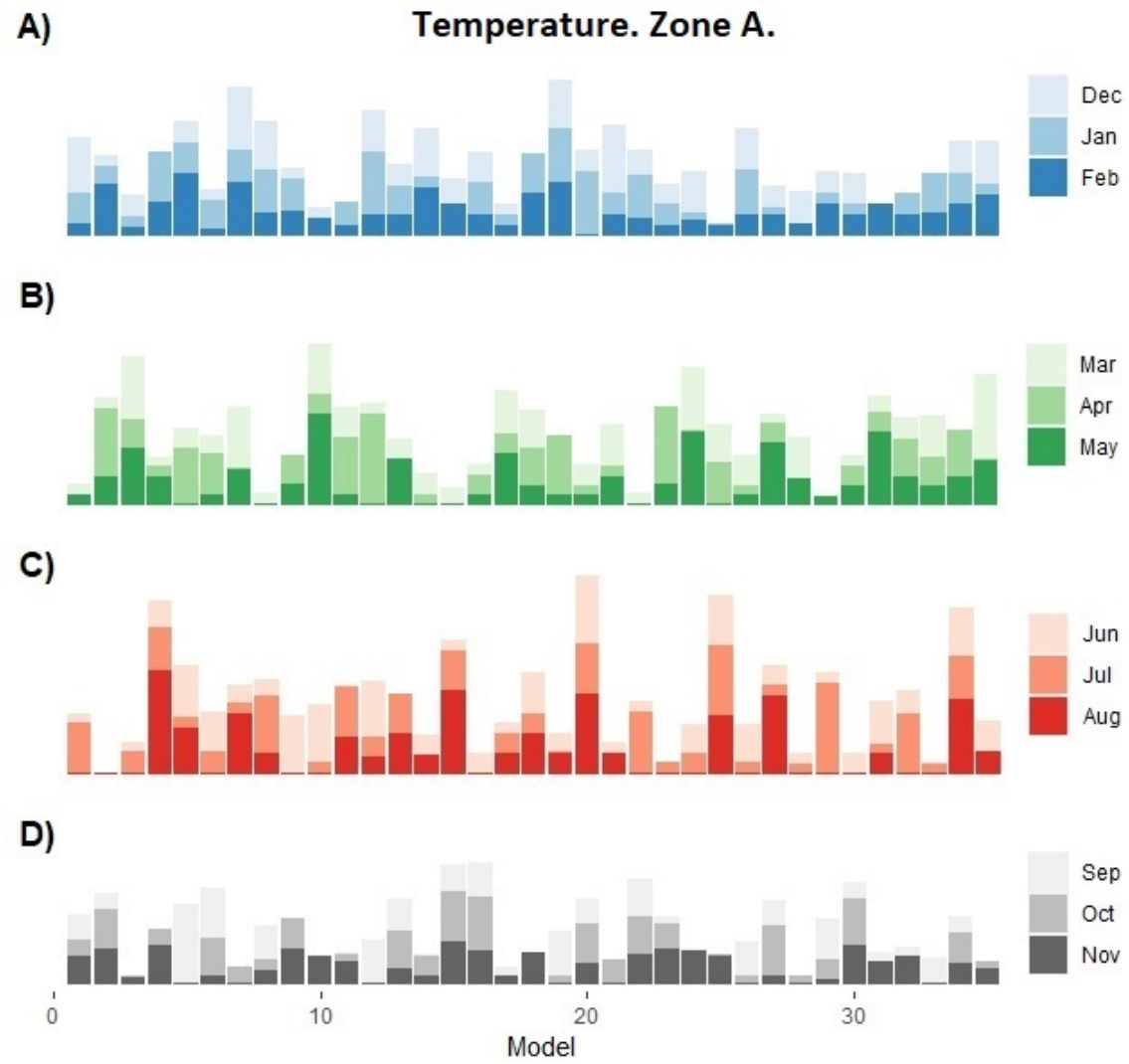

Figure 3.37: VMCE method weights for each month combined per season for temperature data from **climate zone A** - Polar regions (with latitude south from 66.5°S or north from 66.5°N) combined. **A)** December, January and February. **B)** March, April and May. **C)** June, July and August. **D)** September, October and November.

Though there are no models having the same or close weights for all the seasons, the summer season has a distinct subset of models having high weights for all summer months. As the summer season is the time of the year with the least temperature variation in polar

climate zone this indicates the VMCE tends to weight the models on equal (or close) datasets equally.

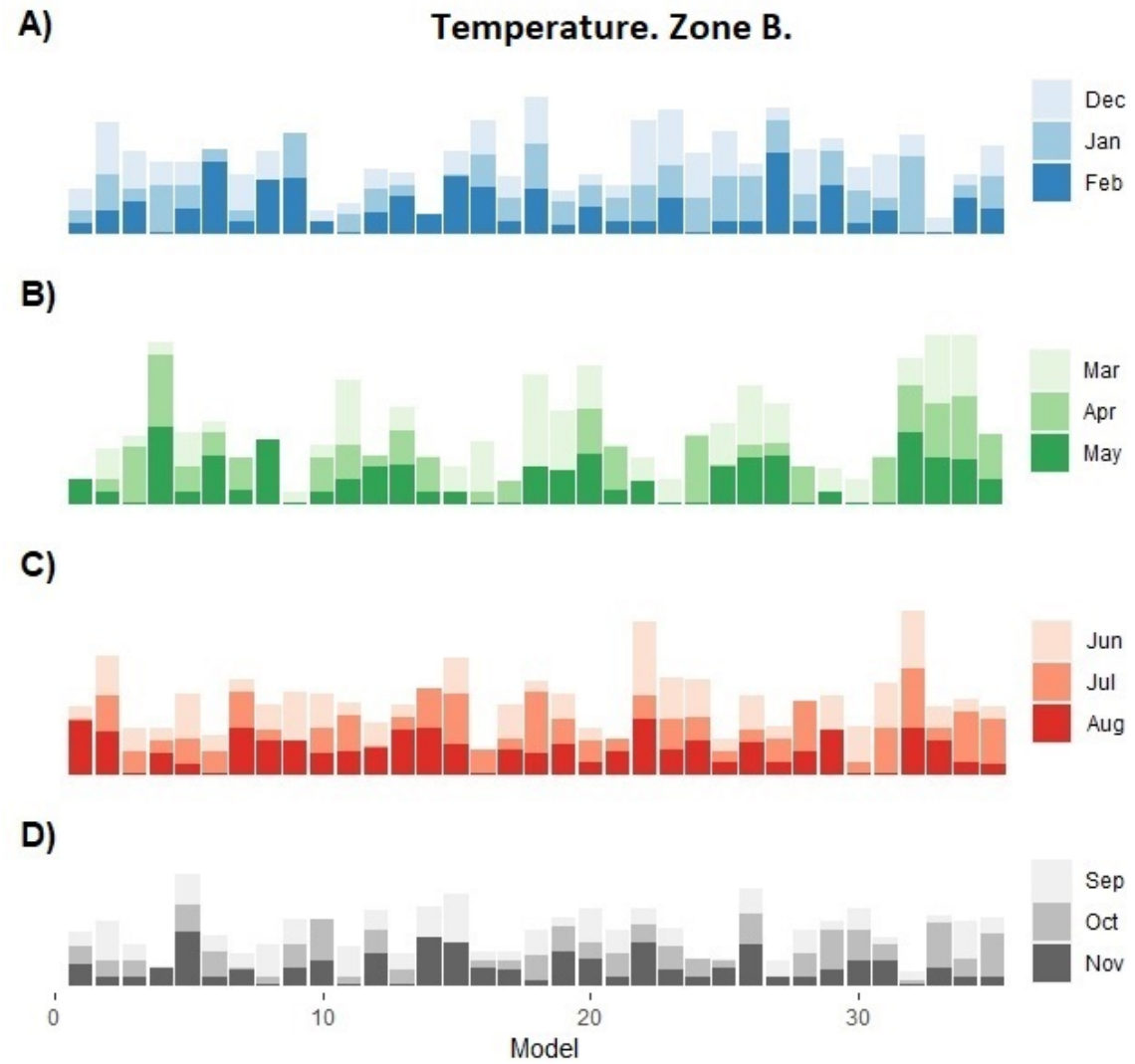

Figure 3.38: VMCE method weights for each month combined per season for temperature data from **climate zone B** - Southern regions between polar and tropical (with latitude between 66.5°S and 23.5°S). **A)** December, January and February. **B)** March, April and May. **C)** June, July and August. **D)** September, October and November.

As the southern regions between polar and tropical has a smaller variation of temperature between months compared to the polar regions (see Figure 3.3) the variation of weights in each season is slightly higher than in polar regions.

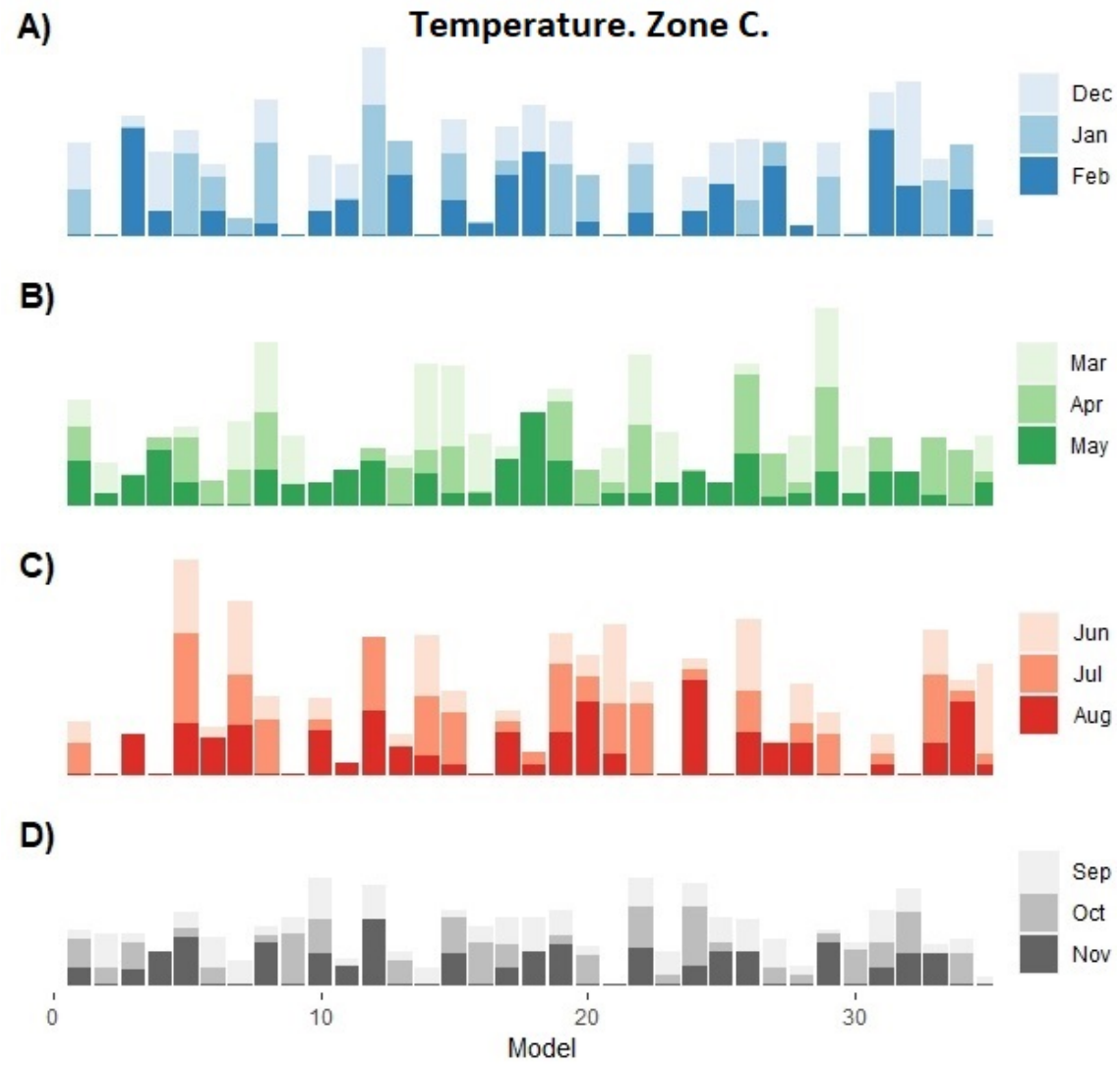

Figure 3.39: VMCE method weights for each month combined per season for temperature data from **climate zone C** - Tropical regions (with latitude between 23.5°S and 23.5°N). **A)** December, January and February. **B)** March, April and May. **C)** June, July and August. **D)** September, October and November.

The variation of weights in each season for tropical regions confirms it has a reverse correlation to the temperature variation.

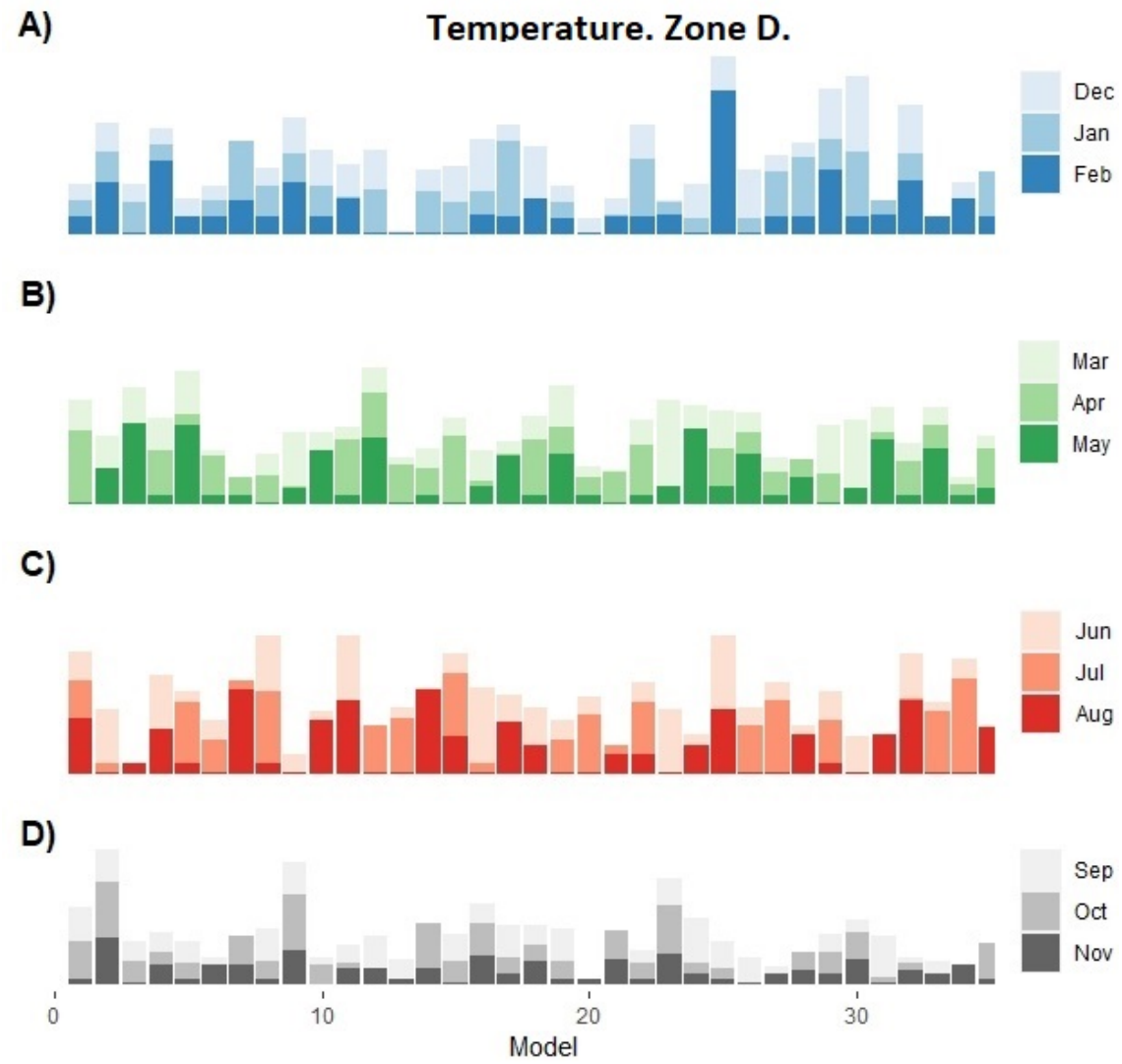

Figure 3.40: VMCE method weights for each month combined per season for temperature data from **climate zone D** - Northern regions between tropical and polar (with latitude between 23.5°N and 66.5°N) **A)** December, January and February. **B)** March, April and May. **C)** June, July and August. **D)** September, October and November.

The weights for the northern regions between tropical and polar are similar in variation to the weights in southern regions between tropical and polar.

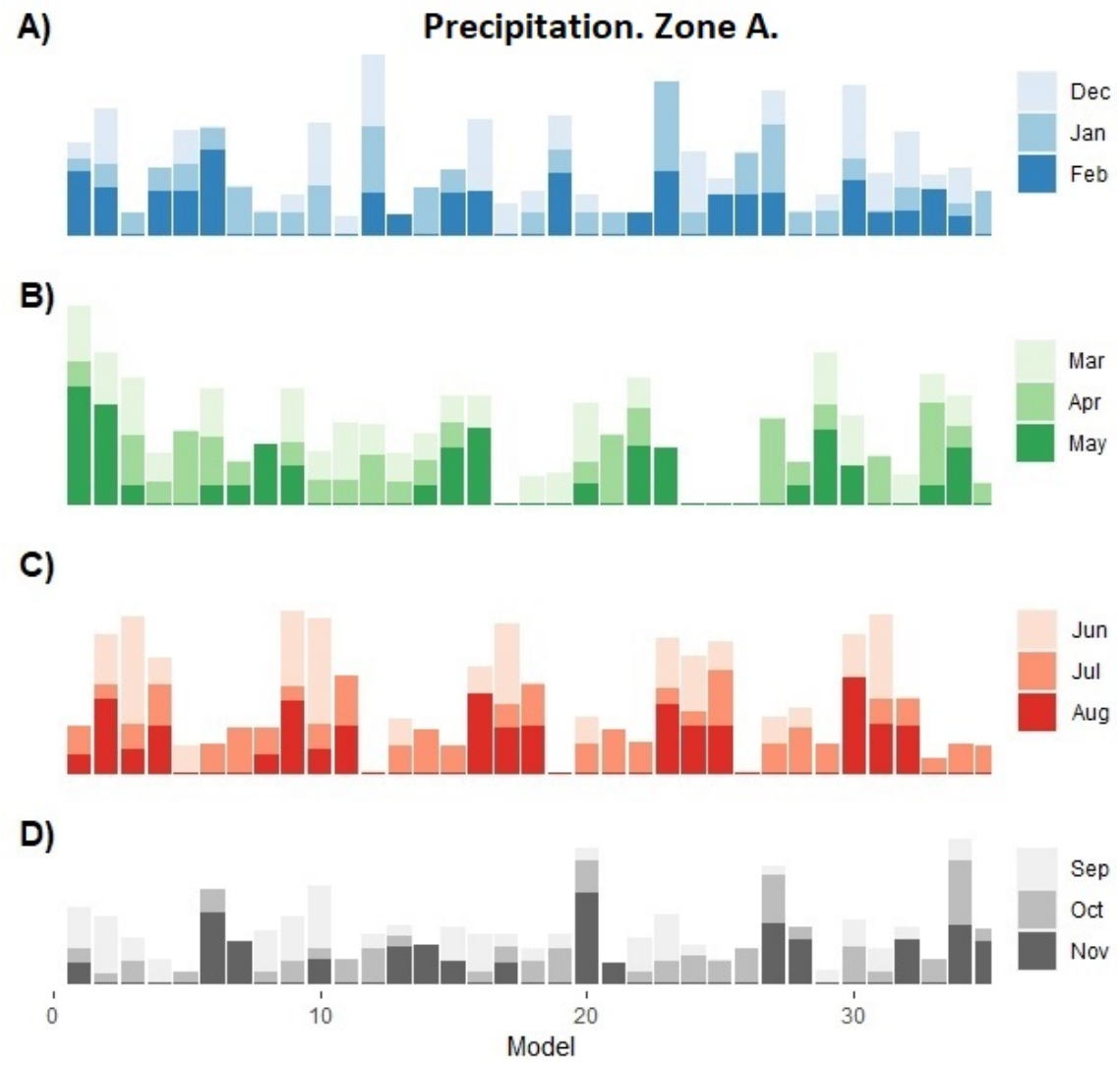

Figure 3.41: VMCE method weights for each month combined per season for precipitation data from **climate zone A** - Polar regions (southern and northern) combined (with latitude south from 66.5°S or north from 66.5°N) combined. **A)** December, January and February. **B)** March, April and May. **C)** June, July and August. **D)** September, October and November.

As the polar regions have the smallest variation of precipitation the weights variation is high within each season. This confirms the findings for temperature data.

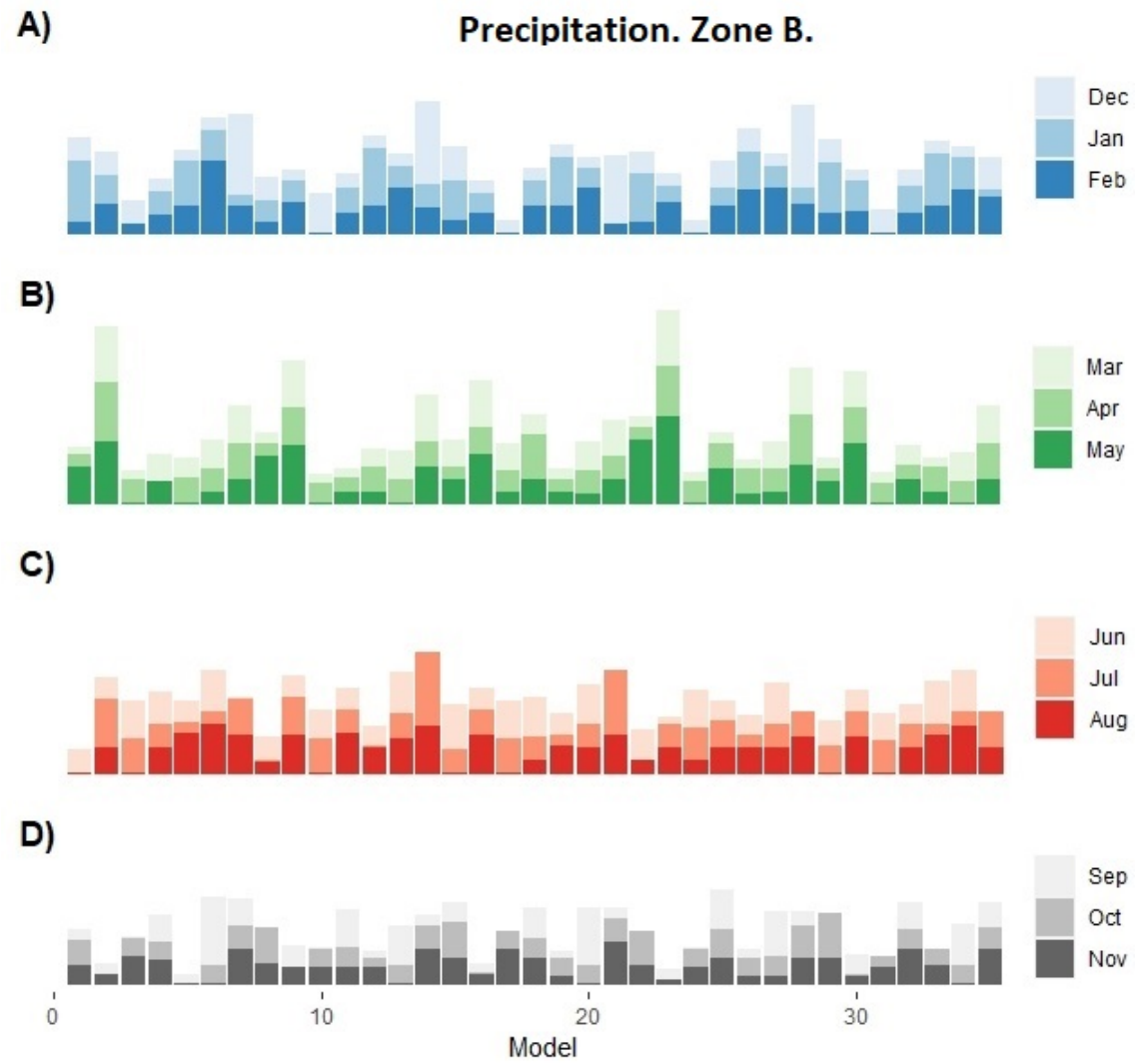

Figure 3.42: VMCE method weights for each month combined per season for precipitation data from **climate zone B** - Southern regions between polar and tropical (with latitude between 66.5°S and 23.5°S). **A)** December, January and February. **B)** March, April and May. **C)** June, July and August. **D)** September, October and November.

The variation between weights within each season is significantly lower for precipitation in southern regions between polar and tropical compared to polar regions.

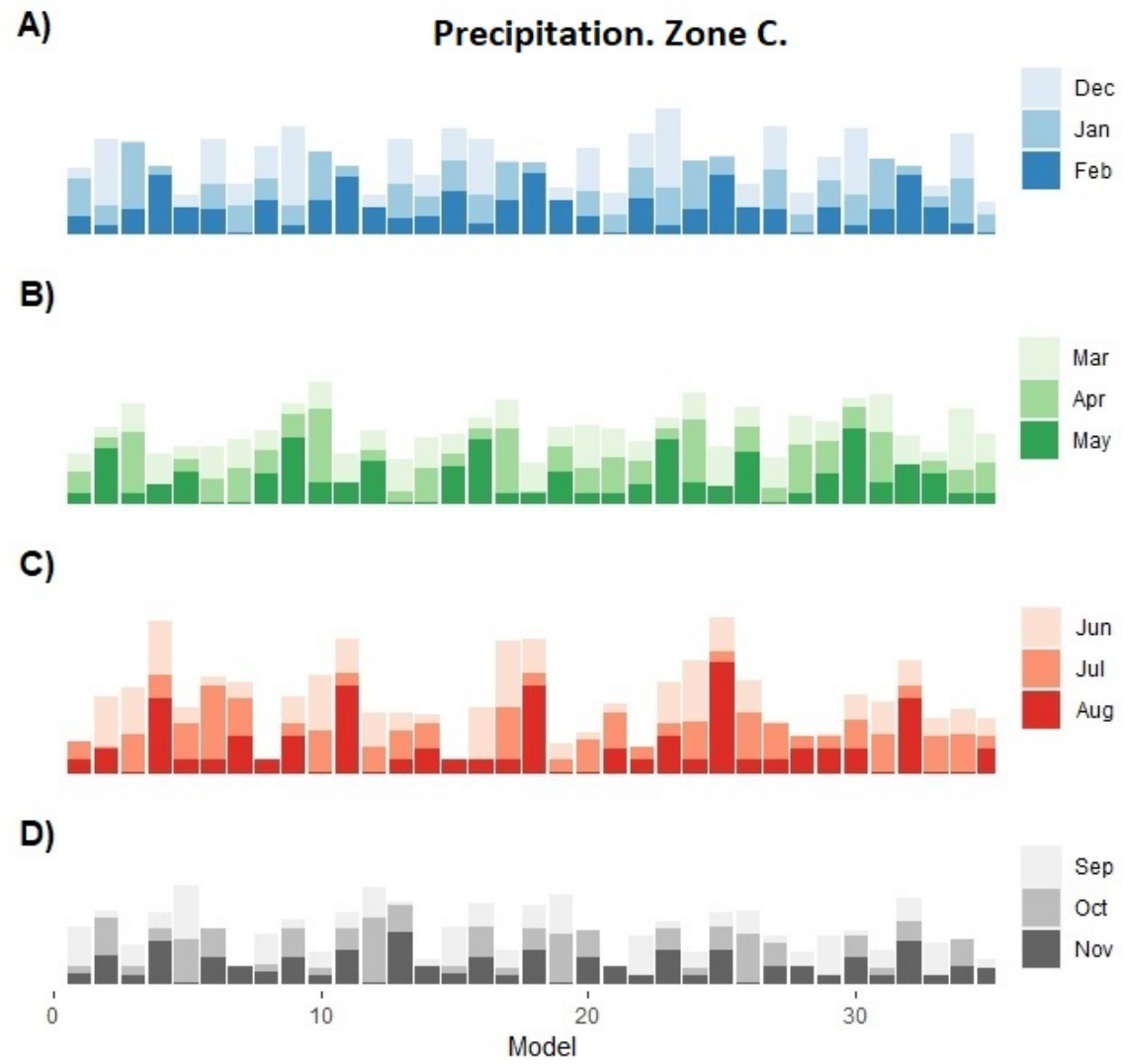

Figure 3.43: VMCE method weights for each month combined per season for precipitation data from **climate zone C** - Tropical regions (with latitude between 23.5°S and 23.5°N). **A)** December, January and February. **B)** March, April and May. **C)** June, July and August. **D)** September, October and November.

The variation between weights for tropical regions is overall the closest to normal compared to other climate zones.

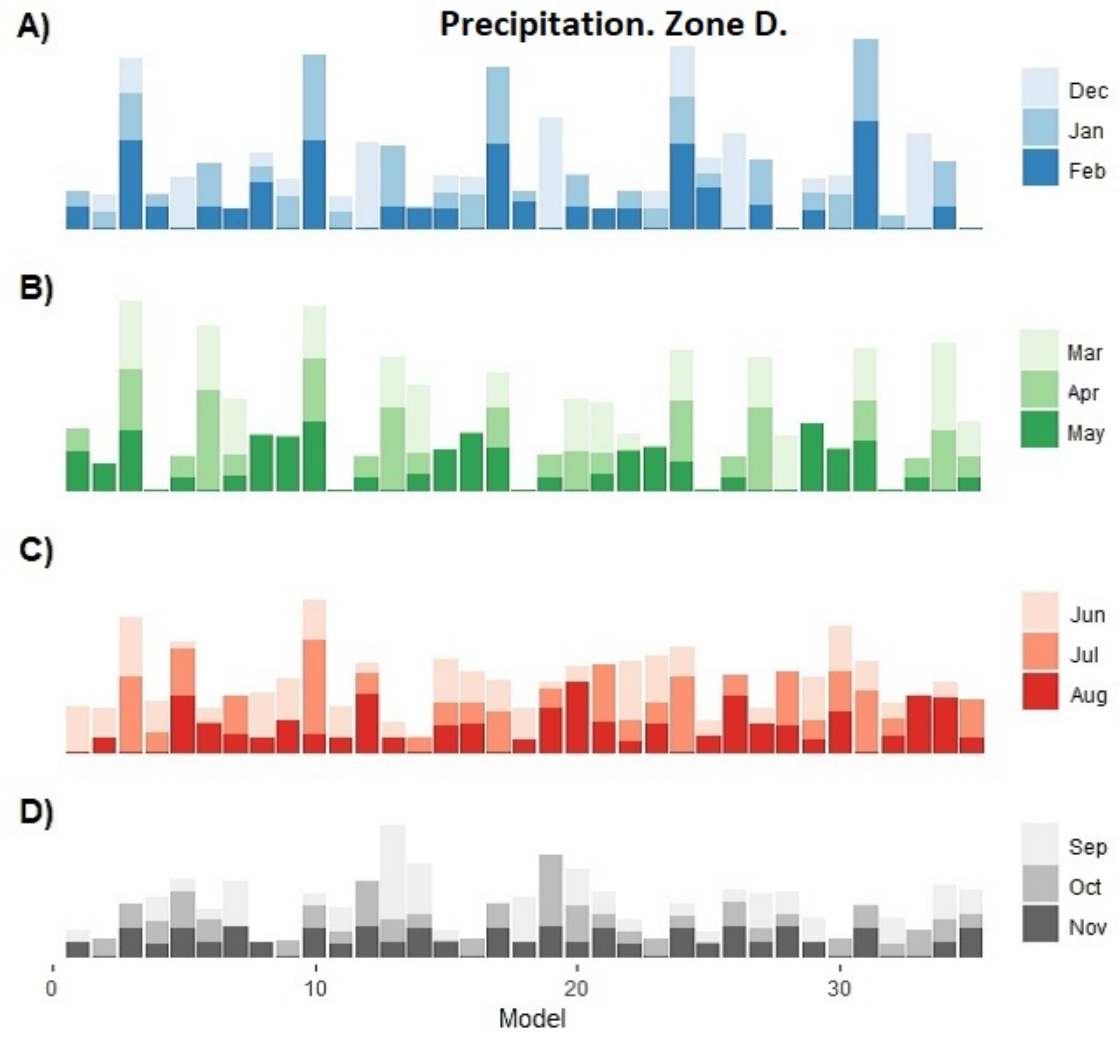

Figure 3.44: VMCE method weights for each month combined per season for precipitation data from **climate zone D** - Northern regions between tropical and polar (with latitude between 23.5°N and 66.5°N). **A)** December, January and February. **B)** March, April and May. **C)** June, July and August. **D)** September, October and November.

The variation between weights for northern regions between tropical and polar is similar to southern regions except for winter and spring seasons, when precipitation is generally more uniform.

The results of the varying weights analysis are indicative as they can differ from location to location even within the same climate zone. By analysing the variation of weights within

each season we were able to find some evidence that this variation is reverse correlated with the variation of the data, which in its turn indicates that the VMCE tends to weight models in a similar way on data with low variability.

3.4 Discussion

We demonstrated that the nonlinear approach (VMCE) to weighting climate ensembles used here has better performance than other methods in terms of climatological metrics. We confirmed the findings in Chapter 2 by applying both MCE and VMCE on a much larger set of data, exposing all the methods to highly divergent data configurations. The results were robust across the spatial grid and climate variables (temperature and precipitation).

To study the effectiveness of varying weights application separately from the effectiveness of Markov chain application, we compared the outcomes from COE and VCOE , MCE and VMCE, COE and MCE, VCOE and VMCE. Varying weights improved performance for both VCOE and VMCE when compared to COE and MCE respectively. The performance improvement is most evident for VMCE on non-Gaussian distributed data with low correlations between ensemble models.

The inherit limitation of varying weighting approach is inability to extract one general set of weights for model comparison and other similar purposes. In this way, the application of VMCE method is best suited for forecast and future climate analysis.

Another inherit limitation of both MCE and VMCE methods comes from the way the distance matrix is constructed (see Section 2.2.2 Markov chain ensemble (MCE) method). As equation 2.2 is based on Euclidean distance between models outputs and observations, the stochastic simulations produced by the MCE and VMCE methods are inevitably concentrated around values closest to observations. This phenomena prevents the MCE and VMCE methods from generating simulations optimal for climatological metrics that are not related to the ensemble proximity to the observations (e.g. trend bias and interan-

nual variability). This limitation can be overcome by replacing equation 2.2 (Euclidean distance criteria) with a criteria which is more suitable to the climatological metrics in focus. Such improvements however require major modifications of the MCE and VMCE methods and are outside of this study's scope.

To demonstrate the potential of this nonlinear approach we selected a simple monthly-based way of splitting weights. It is objective and has a sound motivation that each month has a different mean, variance and other parameters. A more advanced approach to splitting weights might produce much better results than the current implementation of VMCE. Finding an optimal approach to varying ensemble weights is outside of the scope of this study and can be a topic for future research.

To further strengthen robustness and reliability of the proposed Markov Chain - based approach a structured way of measuring uncertainty of the results is required. This will be discussed in Chapter 4 in details.

3.5 Conclusion

To verify and extend the results obtained in Chapter 2 we applied the VMCE method on spatially explicit CMIP6 data optimising its performance by introducing a novel way of constructing a weighted ensemble mean with varying weights. The CMIP6 data in this study contains temperature and precipitation variables in different climate zones, allowing application of the VMCE method on structurally diverse inputs.

The results obtained in Chapter 3 indicate an overall better performance of VMCE method in climatological monthly RMSE and monthly means error, similar level of performance in monthly trend error and lower performance in inter-annual variability error when compared to convex optimisation.

Based on the above, we can conclude that the VMCE method can be applied on different types of data (including spatially explicit data) with relatively high performance. It is

best suited for forecast and future climate analysis.

The MCE and VMCE methods' robustness and reliability can be further strengthened by estimating uncertainty of its results, which will be discussed in details in the following Chapter 4.

Further study can include major modification of the simulations generation step to increase its performance in a wider range of climatological metrics.

Chapter 4

Climate model ensemble uncertainty estimation

4.1 Introduction

In Chapters 2 and 3 we demonstrated that the Markov chain method is a flexible method for constructing weighted climate model ensembles with high performance compared with standard approaches on common metrics. In this chapter, we address the important question concerning its suitability for estimation of future climate uncertainty. As Curry J. (2011) pointed out there is a number of approaches aiming to improve understanding and characterisation of climate uncertainty with many of those approaches based on climate models outputs' spread. A common way of measuring uncertainty of multi-model ensemble with equal weights is to take quantile values of models outputs (Knutti et al. (2017)). In a similar manner weighted quantile values can be used to measure uncertainty of weighted climate model ensembles (Brunner et al. (2020)).

To evaluate the effectiveness of using a weighted quantile interval for uncertainty estimation we compare it to a prediction interval commonly used in the regression and forecasting analysis literature (Johnson and Wichern (2002), p.247). In order to satisfy the necessary

conditions for using prediction interval we transform training data into a normally distributed dataset, calculate the upper and lower control limits and use the retrieved results to estimate the prediction interval of the weighted climate model ensembles.

We apply the prediction interval (PI) and the weighted quantile interval (WQ) on geospatial CMIP6 data for five ensemble weighting methods (AVE, COE, MCE, VCOE and VMCE) described in Chapters 2 and 3. We evaluate the performance of those uncertainty estimations by calculating uncertainty area (UA) and uncertainty error (UE) metrics on training and validation periods for temperature and precipitation variables on CMIP6 data described in Chapter 2. We conclude this chapter by constructing uncertainty estimation intervals using prediction and weighted quantile intervals for global yearly data for both temperature and precipitation data.

4.2 Methods

4.2.1 Data

Here we use the same CMIP6 data as in Chapter 3 (Section 3.2.1).

4.2.2 Prediction interval

Prediction interval is an estimate of an interval in which future observations fall with a certain probability, assuming that observations are normally distributed. It is defined by upper (ULC) and lower (LCL) control limits calculated according to Equation 4.1 adapted from Johnson and Wichern (2002):

$$
\begin{aligned}
UCL_X^{\alpha} &= \overline{X} + T_{\alpha/2,n-1} s_n \sqrt{1+1/n}, \\
LCL_X^{\alpha} &= \overline{X} - T_{\alpha/2,n-1} s_n \sqrt{1+1/n}, \\
s_n^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2
\end{aligned}
\tag{4.1}
$$

where X is a data sample of size n, $\overline{X}$ is the mean value of X and $T_{\alpha/2,n-1}$ is $100(1-\alpha/2)$th percentile of Student's t-distribution.

We modify this prediction interval calculation to reflect the non-Gaussian distribution of the climate data (see Figures 3.3 and 3.4) by applying it on residuals of climate approximation by weighted ensemble mean. The detailed algorithm for the modified prediction interval calculation is described in Table 4.1 below:

Notation:

- O_t is the observation value at time t
- E_t is the weighted ensemble mean value at time t
- T_1 is the length of training period
- T is the length of uncertainty estimation time period
- ε_t is the residual between observation and weighted ensemble values at time t
- σ_ε is the standard deviation of ε
- $F_{\alpha,m-1}$ is the upper $\alpha \times 100\%$ point of the Student distribution with $m-1$ degrees of freedom (one-sided).

Step 1. Calculate residuals $\varepsilon_t = O_t - E_t$ for each $t \in [1, T_1]$

Step 2. If $\varepsilon_t \geq 0$ for each $t \in [1, T_1]$, set $LCL_t^{1-\alpha}(PI) = E_t$ and go to Step 5

Step 3. Construct a subset $\varepsilon^L = \varepsilon < 0$ of length T_L

Step 4. Calculate the lower control limit value according to equation 4.2

$$LCL_t^{1-\alpha}(PI) = E_t + \overline{\varepsilon^L} - F_{\alpha,T_L-1}\sigma_\varepsilon^L\sqrt{1+1/T_L} \tag{4.2}$$

Step 5. If $\varepsilon_t \leq 0$ for each $t \in [1, T_1]$, set $UCL_t^{1-\alpha}(PI) = E_t$ and go to Step 8

Step 6. Construct a subset $\varepsilon^U = \varepsilon > 0$ of length T_U

Step 7. Calculate he upper control limit value according to equation 4.3

$$UCL_t^{1-\alpha}(PI) = E_t + \overline{\varepsilon^U} + F_{\alpha,T_U-1}\sigma_\varepsilon^U\sqrt{1+1/T_U} \tag{4.3}$$

Step 8. Combine lower and upper control limit values to construct the prediction interval $PI^{1-\alpha} = [LCL_t^{1-\alpha}(PI), UCL_t^{1-\alpha}(PI)]$

Step 9. Repeat steps 1-8 for each $t \in [1, T]$

Table 4.1: Prediction interval (PI) algorithm.

The procedure can be justified as follows. It makes sense to model the residuals distribution as a mixture of two distributions: of positive residuals (distribution A) and of negative residuals (distribution B). When a particular ε_t is realised, it is clearly from A (if it is positive) or from B (if it is negative). Then, a $(1-\alpha) \times 100\%$ prediction interval is constructed conditionally on this information and the interval is superimposed on E_t. The

resulting interval is asymmetric and includes the weighted ensemble mean value E. It is not directly dependent on climate model values or climate model ensemble weights and can be estimated for the future periods using only the weighted ensemble mean values for those periods. Those properties of prediction interval simplify its calculation, interpretation and analysis while allowing its application on any type of data where the residuals distributions are close to normal. We present the distributions of residuals $\varepsilon < 0$ and $\varepsilon > 0$ for temperature and precipitation in Figure 4.1 below.

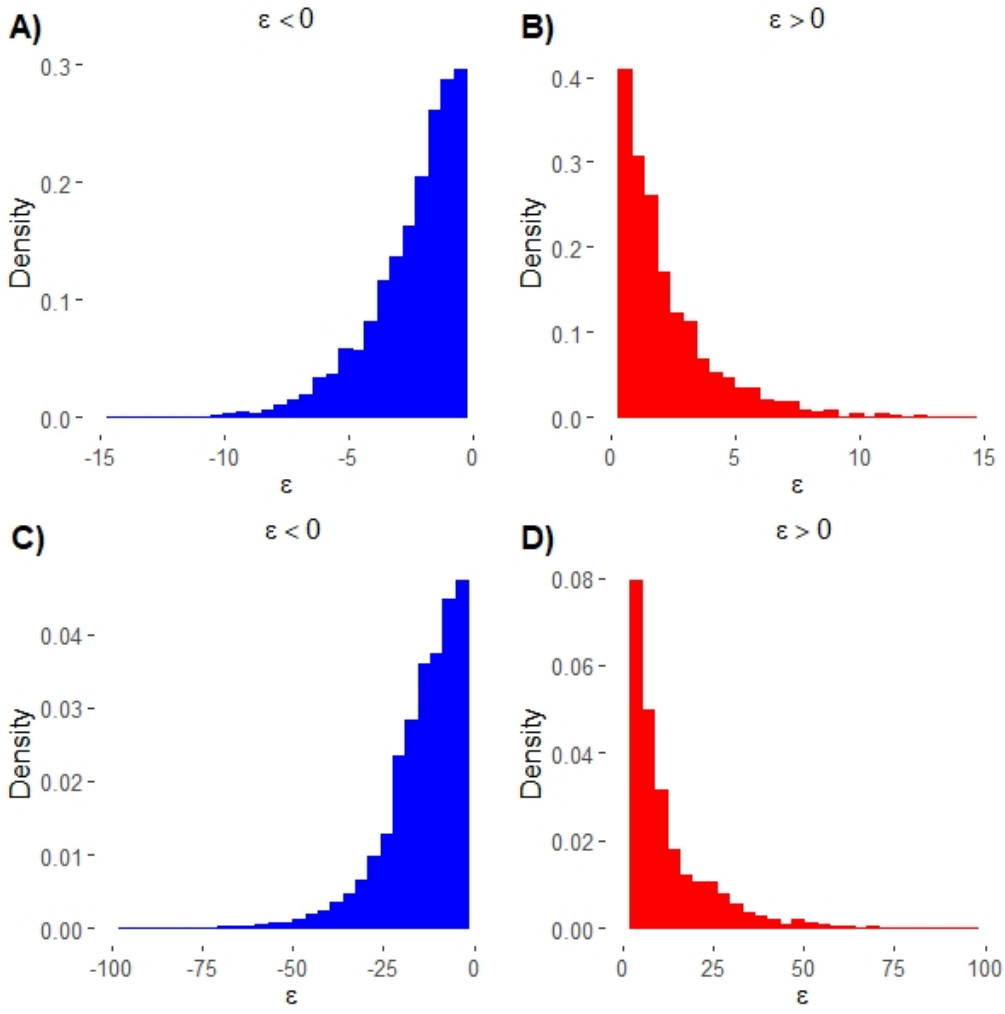

Figure 4.1: Residuals' distribution on climate data for AVE, COE, MCE, VCOE and VMCE combined. **A)** Histogram showing negative residuals' distribution for temperature. **B)** Histogram showing positive residuals' distribution for temperature. **C)** Histogram showing negative residuals' distribution for precipitation. **D)** Histogram showing positive residuals' distribution for precipitation.

4.2.3 Weighted quantile

The weighted quantile (WQ) interval is based on the climate model output values and the climate model ensemble weights. At each time point t model outputs are sorted in ascending order. Their weights are sorted in the same order and their corresponding cumulative sums are calculated. If a cumulative sum of the first x weights is equal $1-\alpha/2$ an average of this sum and the sum of $x+1$ weights is taken as the weighted quantile value. If a cumulative sum of the first x weights is bigger than the $1-\alpha/2$ value this sum is taken as the weighted quantile value. The detailed algorithm for the weighted quantile interval calculation is described in Table 4.2 below:

Notation:

- N is the number of models in ensemble
- M_t^i is the i model value at time t
- w_i is the ensemble weight of the i model
- T is the length of uncertainty estimation time period

Step 1. Sort N models' outputs in ascending order: $M_t^j \leq M_t^{j+1} \leq ... \leq, M_t^J$ where $(j, j+1, ..., J) \in (i, i+1, ..., N)$

Step 2. Sort N models' weights according to Step 1 results: $(w_j, w_{j+1}, ..., w_J)$

Step 3. Calculate cumulative weights sums: $\sum_{k=1}^{1} w_k, \sum_{k=1}^{2} w_k, ..., \sum_{k=1}^{N} w_k$

Step 4. If $\sum_{k=1}^{x} w_k = \alpha/2$ the lower control limit value $LCL_t^{1-\alpha}(WQ) = (M_t^x + M_t^{x+1})/2$

Step 5 If $\sum_{k=1}^{x} w_k < \alpha/2 < \sum_{k=1}^{x+1} w_k$, the lower control limit value $LCL_t^{1-\alpha}(WQ) = M_t^{x+1}$.

Step 6. If $\sum_{k=1}^{x} w_k = 1-\alpha/2$, upper control limit value $UCL_t^{1-\alpha}(WQ) = (M_t^x + M_t^{x+1})/2$

Step 7 If $\sum_{k=1}^{x} w_k < 1-\alpha/2 < \sum_{k=1}^{x+1} w_k$ upper control limit value $UCL_t^{1-\alpha}(WQ) = M_t^{x+1}$

Step 8. Combine upper and lower control limit values to construct the weighted quantile interval $WQ^{1-\alpha} = [LCL^{1-\alpha}(WQ), UCL^{1-\alpha}(WQ)]$

Step 9. Repeat Steps 1-8 for each $t \in [1, T]$

Table 4.2: Weighted quantile (WQ) algorithm.

The resulting interval is asymmetric and doesn't require calibration on training period. It can be estimated using ensemble models' outputs and a corresponding set of weights. It is not dependent on the ensemble performance in terms of proximity to observations and it does not require any assumptions on data distribution. Those properties of weighted quantile interval allow its application on any type of data, however its performance is heavily dependent on the performance and distribution of the included models as we will demonstrate below.

4.2.4 Performance metrics

To evaluate effectiveness of uncertainty estimation for different ensemble weighting methods (AVE, COE, MCE, VCOE and VMCE) we calculate uncertainty area (UA) as the average difference between upper (UCL) and lower (LCL) control limits and uncertainty error (UE) as a difference between a proportion of observations within control limits (LCL and UCL) and the respective $1-\alpha$ value.

Uncertainty area is calculated as an average difference between upper and lower control limits for each grid cell for a selected period T according to Equation 4.4:

$$UA^{1-\alpha} = \frac{\sum_{t=1}^{T} UCL_t^{1-\alpha} - LCL_t^{1-\alpha}}{T} \tag{4.4}$$

Uncertainty error is calculated as a difference between a proportion of observations within control limits ($LCL^{1-\alpha}$ and $UCL^{1-\alpha}$) and the respective $1-\alpha$ value according to Equation 4.5.

$$UE^{1-\alpha} = \frac{\sum_{t=1}^{T} I(O_t)}{T} - (1-\alpha), I(O_t) = \begin{cases} 1, \text{if } LCL_t^{1-\alpha} \leq O_t \leq UCL_t^{1-\alpha} \\ 0, \text{if } LCL_t^{1-\alpha} > O_t \text{ or } UCL_t^{1-\alpha} < O_t \end{cases} \tag{4.5}$$

4.2.5 Cross-validation procedures

Here we employ the same cross-validation procedures as in Chapter 2 and Chapter 3 to assess how the uncertainty estimation results obtained on calibration data will perform on new independent datasets.

Holdout method is described in detail in Chapter 2 (Section 2.2.6.1) and is based on splitting the dataset into training and validation sets. The goal of cross-validation is to examine the model's ability to predict new data that was not used in estimating the required parameters.

Model-as-truth performance assessment is described in detail in Chapter 2 (Section 2.2.6.2) and is based on selecting one model as a true model (pseudo-observations) with the remaining models used to build a weighted ensemble mean that best estimates the true model over the historical period. This weighted ensemble mean is then tested against the future projections of the true model. This procedure is repeated N times with each of the ensemble members being a true model where N is the total number of the ensemble members.

4.3 Results

We summarise the UA and UE results in the following tables as global aggregated means of all land points weighted by the grid cell sizes for prediction and weighted quantile intervals together with detailed statistical and spatial distributions of those metrics. For consistency in the figures and tables below we use the following notation:

$$[Metric\,abbreviation]^{[1-\alpha\ value]}_{[Time\ period]}([Uncertainty\ estimation\ method]_{[Climate\ variable\ name]})$$

with metric abbreviation being UE for uncertainty error and UA for uncertainty area; time period being Tr for training period (years 1901-1980), V for validation period (years 1981-2020) and $M-a-T$ for model-as-truth experiment period (years 2021-2100); $1-\alpha$ value being equal to 0.65, 0.7, 0.75, 0.8, 0.85, 0.90 or 0.95; uncertainty estimation being PI for prediction interval and WQ for weighed quantile interval; and climate variable name being $Temp.$ for temperature and $Precip.$ for precipitation.

4.3.1 Temperature data

4.3.1.1 Prediction interval for temperature data

We present the aggregated values of prediction interval uncertainty error ($UE^{1-\alpha}(PI_{Temp.})$) on temperature data using all land points weighed according to the grid cells' sizes in Table 4.3 for training period and Table 4.4 for validation period. Those results are accompanied by the prediction interval uncertainty area ($UA^{1-\alpha}(PI_{Temp.})$) results on temperature data averaged over both training and validation period in Figure 4.5.

$1-\alpha$	*0.65*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*
AVE	**0.02**	**0.02**	**0.01**	**0.00**	-0.01	**-0.01**	**-0.01**
COE	0.05	0.04	0.02	0.01	-0.00	**-0.01**	-0.02
MCE	0.04	0.03	**0.01**	**0.00**	-0.01	**-0.01**	-0.02
VCOE	0.07	0.05	0.04	0.02	**0.00**	**-0.01**	-0.02
VMCE	0.07	0.06	0.04	0.02	**0.00**	**-0.01**	-0.03

Table 4.3: Average prediction interval uncertainty error results using all land points weighted according to their area sizes on training period **(years 1901-1980)** for temperature. The smallest errors in each column are emphasised in bold.

$1-\alpha$	*0.65*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*
AVE	0.02	0.01	**0.00**	**-0.01**	**-0.02**	**-0.02**	**-0.02**
COE	0.02	0.01	**0.00**	**-0.01**	-0.03	-0.04	-0.04
MCE	**0.01**	**0.00**	-0.01	-0.02	-0.03	-0.03	-0.03
VCOE	**-0.01**	-0.03	-0.04	-0.05	-0.07	-0.07	-0.08
VMCE	**0.01**	**0.00**	-0.02	-0.03	-0.04	-0.06	-0.06

Table 4.4: Average prediction interval uncertainty error results using all land points weighted according to their area sizes on validation period **(years 1981-2020)** for temperature. The smallest errors in each column are emphasised in bold.

$1-\alpha$	*0.65*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*
AVE	3.56	3.82	4.10	4.42	4.80	5.30	6.07
COE	2.76	2.99	3.24	3.53	3.87	4.32	5.00
MCE	3.27	3.53	3.81	4.13	4.51	5.01	5.77
VCOE	**2.04**	**2.23**	**2.43**	**2.67**	**2.96**	**3.33**	**3.91**
VMCE	2.16	2.36	2.59	2.85	3.16	3.56	4.18

Table 4.5: Average prediction interval uncertainty area results using all land points weighted according to their area sizes on training and validation period **(years 1901-2020)** for temperature. The smallest area sizes in each column are emphasised in bold.

The results for temperature data on training period show that the prediction interval reflects $1-\alpha$ values adequately for all the methods in this study with AVE having a slight advantage. This is confirmed by the results on validation period. All of the methods are relatively close in $UE^{1-\alpha}$ metric, but have a significant difference in $UA^{1-\alpha}$ metric with VCOE method having the most narrow prediction interval.

While not having the best performance in either of the metrics, VMCE has slightly better $UE^{1-\alpha}$ results than VCOE on validation period while having a relatively small difference in $UA^{1-\alpha}$. And it has a significantly lower $UA^{1-\alpha}$ values than AVE which has slightly better $UE^{1-\alpha}$ results on validation period. It indicates that in terms of overall results across both $UE^{1-\alpha}$ and $UA^{1-\alpha}$ metrics, VMCE has a relatively high performance.

The detailed prediction interval ($PI_{Temp.}$) results with $1-\alpha = 0.65, 0.80, 0.95$ are presented as maps of uncertainty errors ($UE^{1-\alpha}$) calculated according to Equation 4.5. The results for training (years 1901-1980) and validation (years 1981-2020) periods are followed by model-as-truth experiment summary for validation period (years 2021-2100). The violin plots describe the distribution of metric values for each method using all land points with white lines showing median values. The spatial plots describe the spatial distribution of metric values for each method with according method names in title. All the plots exclude small portions of data which constitute long tails of the data distributions. The minimum and maximum included values are specified in violin plots' axis and in the captions. Darker colours indicate geographical locations with larger errors.

Figures 4.2, 4.3 and 4.4 show detailed prediction interval uncertainty error results with $1-\alpha = 0.65$ for temperature below.

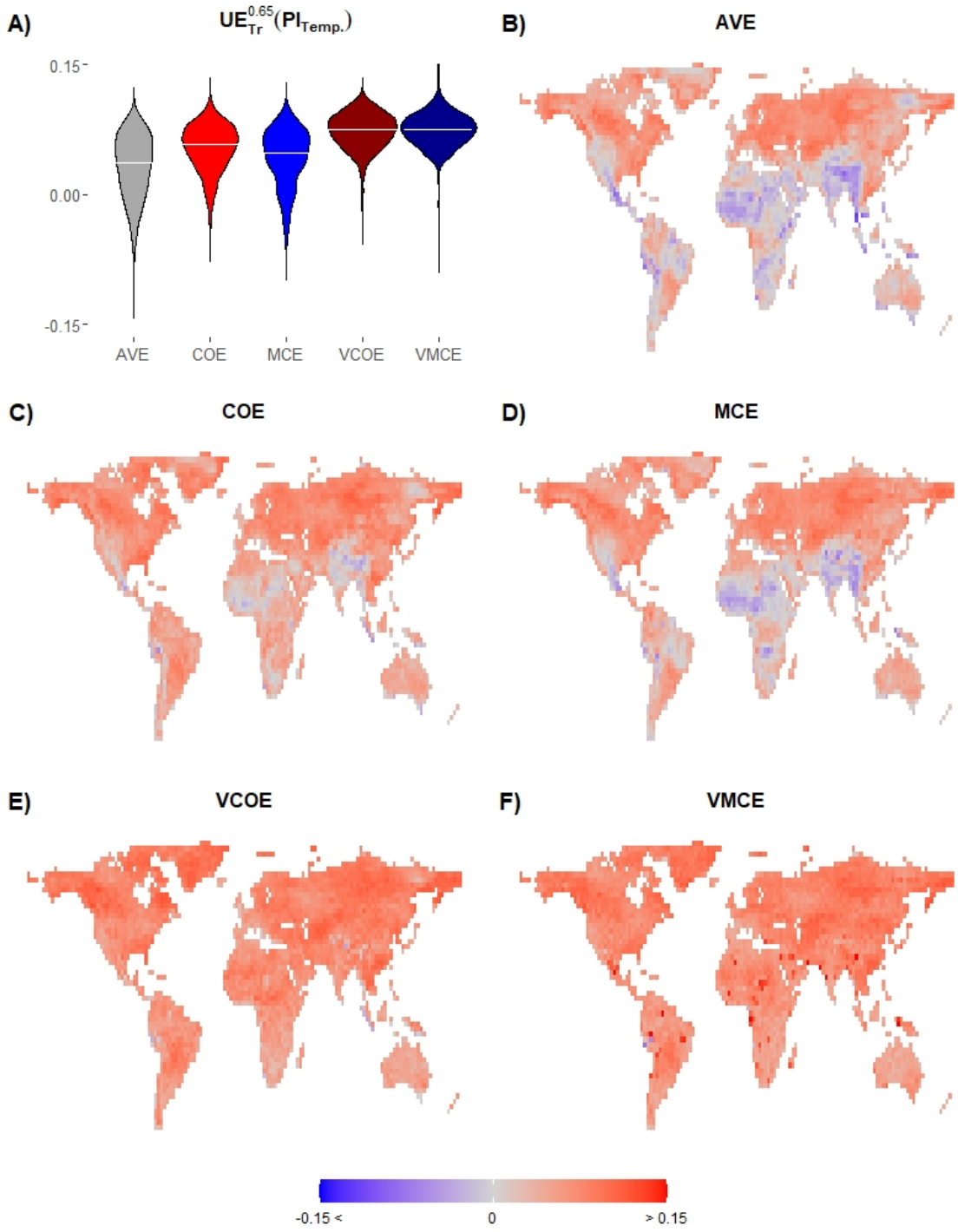

Figure 4.2: Prediction interval uncertainty errors with $1 - \alpha = 0.65$ on training period **(years 1901-1980)** for temperature. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.15 and 0.15. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors respectively.

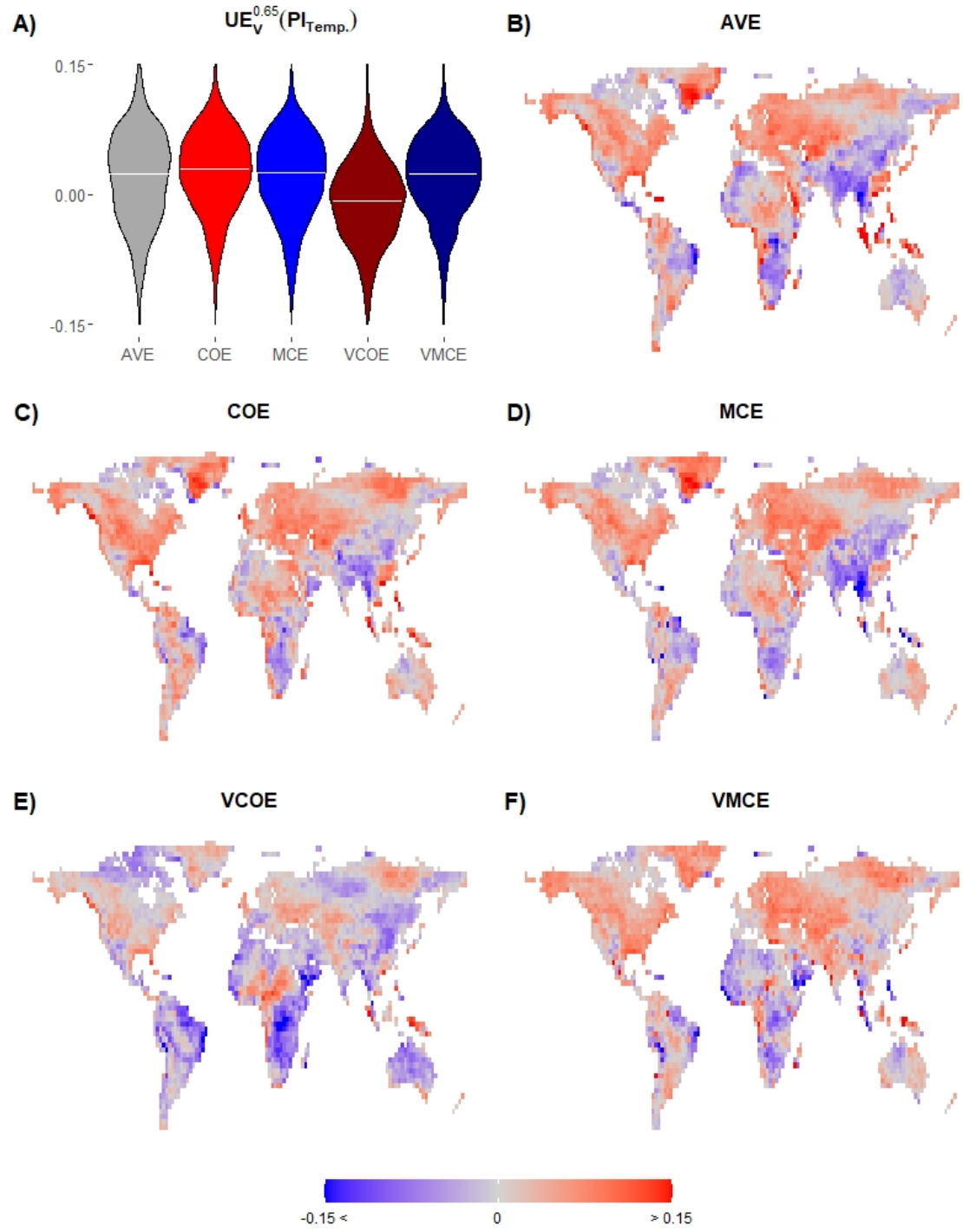

Figure 4.3: Prediction interval uncertainty errors with $1 - \alpha = 0.65$ on validation period **(years 1981-2020)** for temperature. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.15 and 0.15. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors respectively.

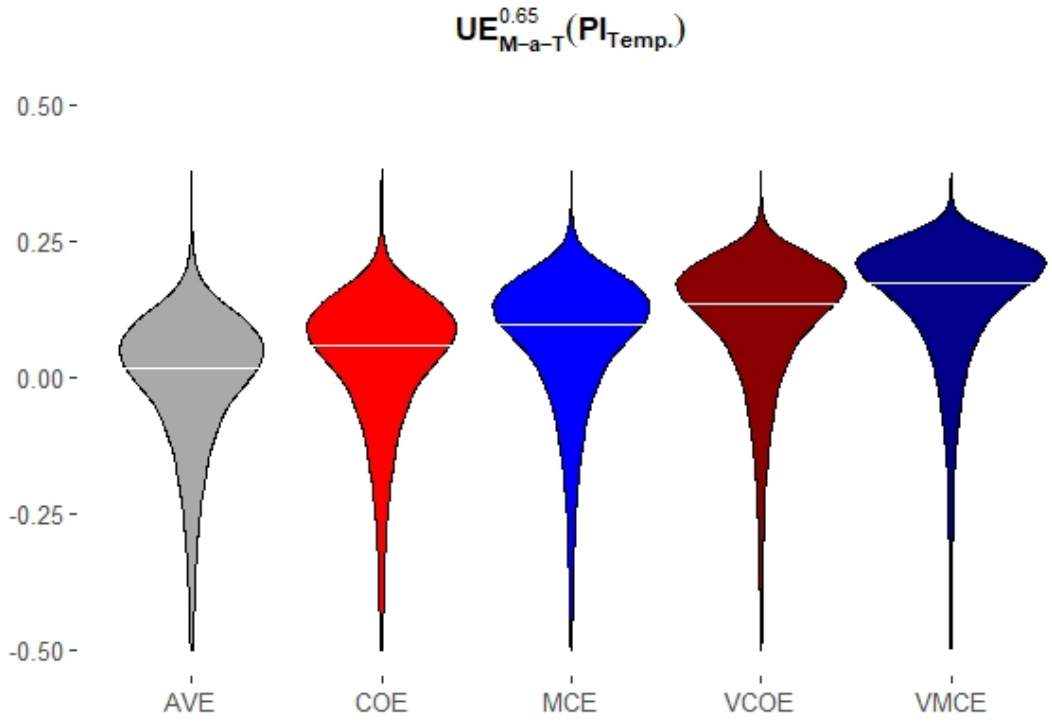

Figure 4.4: Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors with $1 - \alpha = 0.65$ in model-as-truth experiments **(years 2021-2100)** for temperature. The Y-axis is cut at -0.50.

The geo-spatial distributions of uncertainty errors on training period ($UE^{0.65}_{Tr}$) show that the slight advantage of AVE method is due to its $UE^{0.65}_{Tr}$ consisting of a more equal amount of negative and positive values than $UE^{0.65}_{Tr}$ of other methods. This is confirmed by the the geo-spatial distributions of uncertainty errors on validation period ($UE^{0.65}_{V}$). All methods have similar negative and positive $UE^{0.65}_{V}$ distributions on validation period with COE having more negative $UE^{0.65}_{V}$ values (located mostly in Southern hemisphere) than other methods. The model-as-truth experiments uncertainty errors' ($UE^{0.65}_{M-a-T}$) distributions show a slight advantage of AVE over other methods with VMCE having the largest ($UE^{0.65}_{M-a-T}$) median value. As $UE^{0.65}(PI_{Temp.})$ performance is high on validation period but noticeably worse in model-as-truth experiment it indicates that prediction interval is more suitable for near future uncertainty estimation at $1 - \alpha = 0.65$. We compare these $PI^{0.65}$ results with $PI^{0.8}$ results for $1 - \alpha = 0.8$ in Figures 4.5, 4.6 and 4.7.

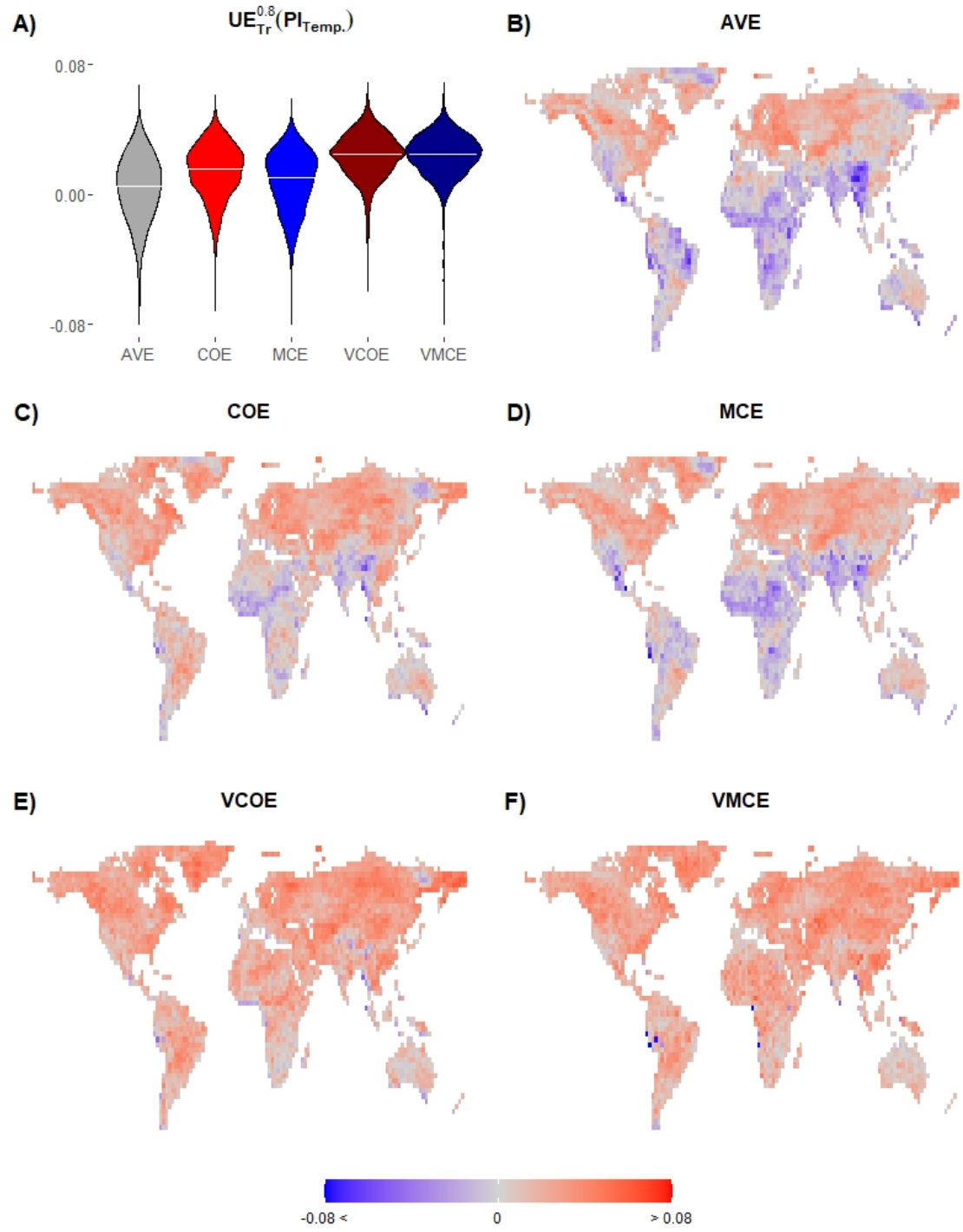

Figure 4.5: Prediction interval uncertainty errors with $1 - \alpha = 0.8$ on training period **(years 1901-1980)** for temperature. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.08. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors respectively.

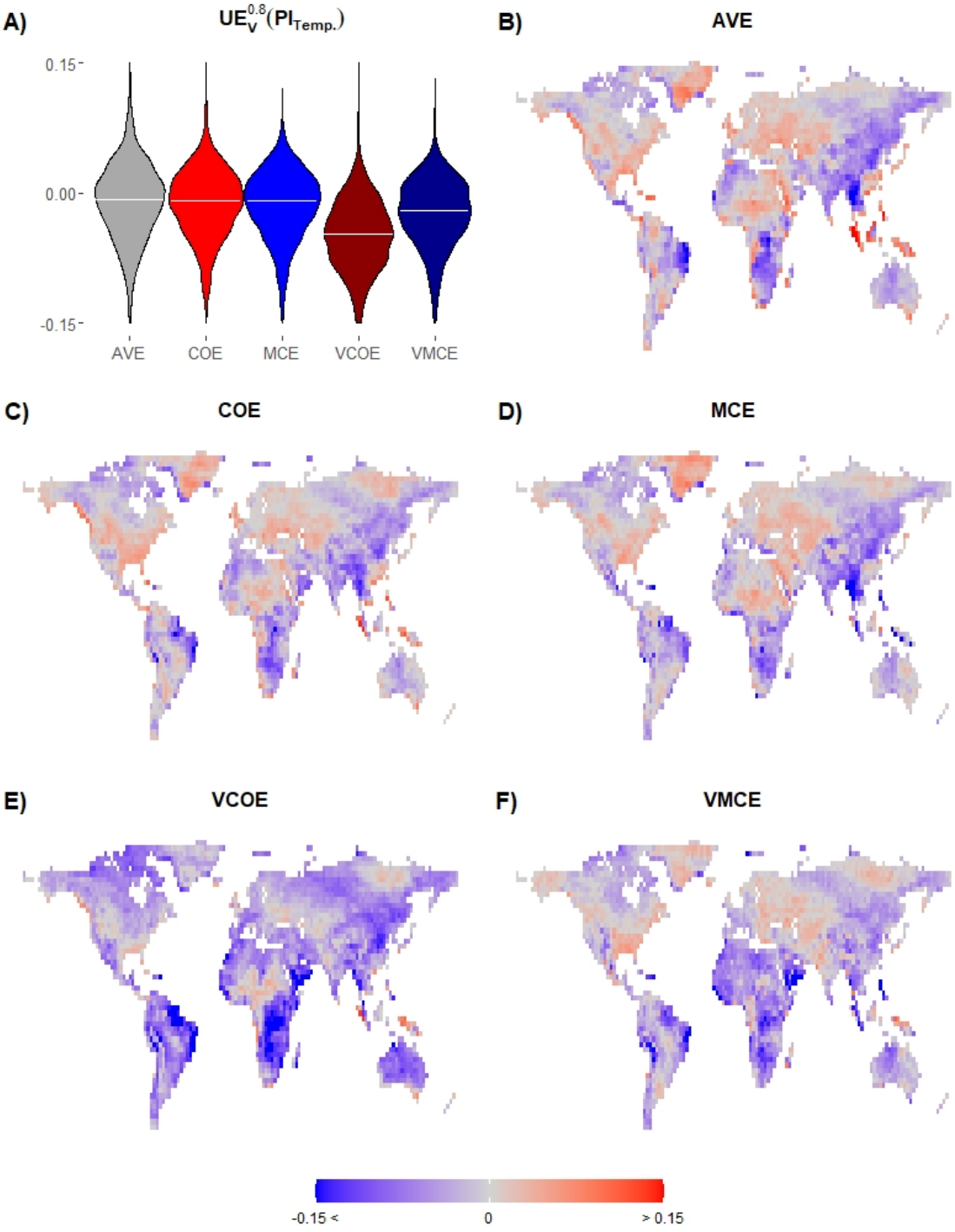

Figure 4.6: Prediction interval uncertainty errors with $1 - \alpha = 0.8$ on validation period **(years 1981-2020)** for temperature. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.15 and 0.15. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors respectively.

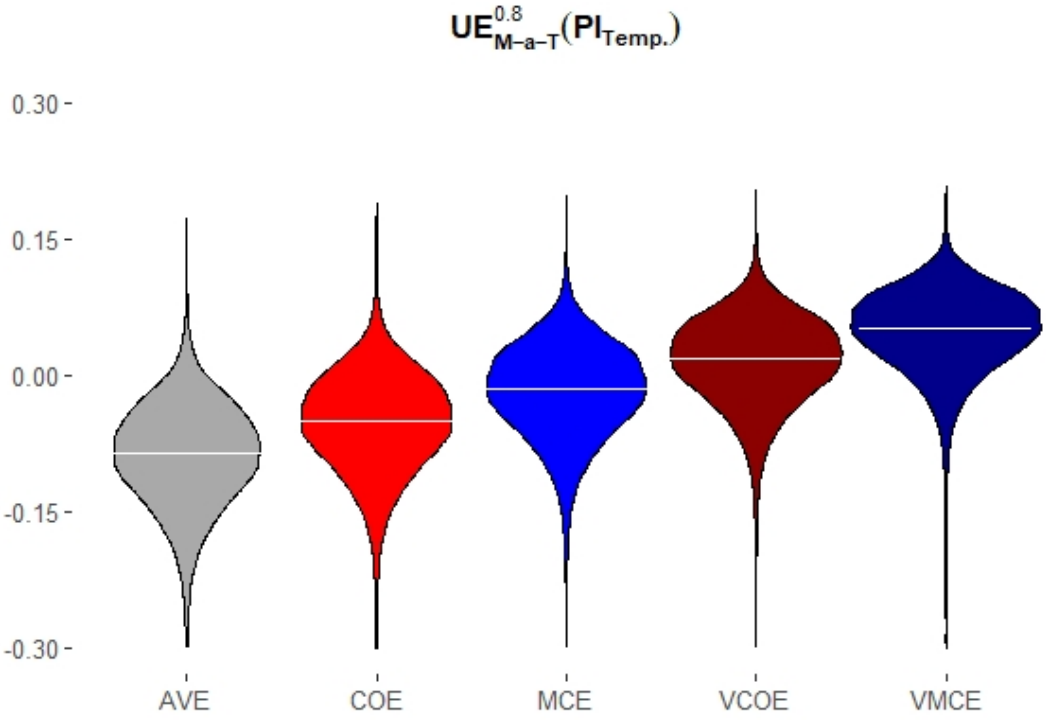

Figure 4.7: Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors with $1-\alpha = 0.8$ in model-as-truth experiments **(years 2021-2100)** for temperature. The Y-axis is cut at -0.30.

The training data shows that $UE^{0.8}_{Tr}$ results are consistently closer to $1-\alpha = 0.8$ across the globe compared to the $UE^{0.65}_{Tr}$ results with $1-\alpha = 0.65$ on training data. This is confirmed by $UE^{0.8}_{V}$ and $UE^{0.8}_{M-a-T}$ results. All methods have similar negative and positive $UE^{0.8}_{V.}$ distributions on validation period with COE having more negative $UE^{0.8}_{V.}$ values (located mostly in Southern hemisphere) than other methods. The model-as-truth experiments uncertainty errors' ($UE^{0.8}_{M-a-T}$) distributions show a slight advantage of MCE over other methods with AVE having the smallest and VMCE having the largest ($UE^{0.8}_{M-a-T}$) median value. As $UE^{0.8}(PI_{Temp.})$ performance is high on validation period but noticeably worse in model-as-truth experiment it indicates that prediction interval is more suitable for near future uncertainty estimation at $1-\alpha = 0.8$. We compare these $PI^{0.8}$ results with $PI^{0.95}$ results for a commonly used $1-\alpha = 0.95$ in Figures 4.8, 4.9 and 4.10 below.

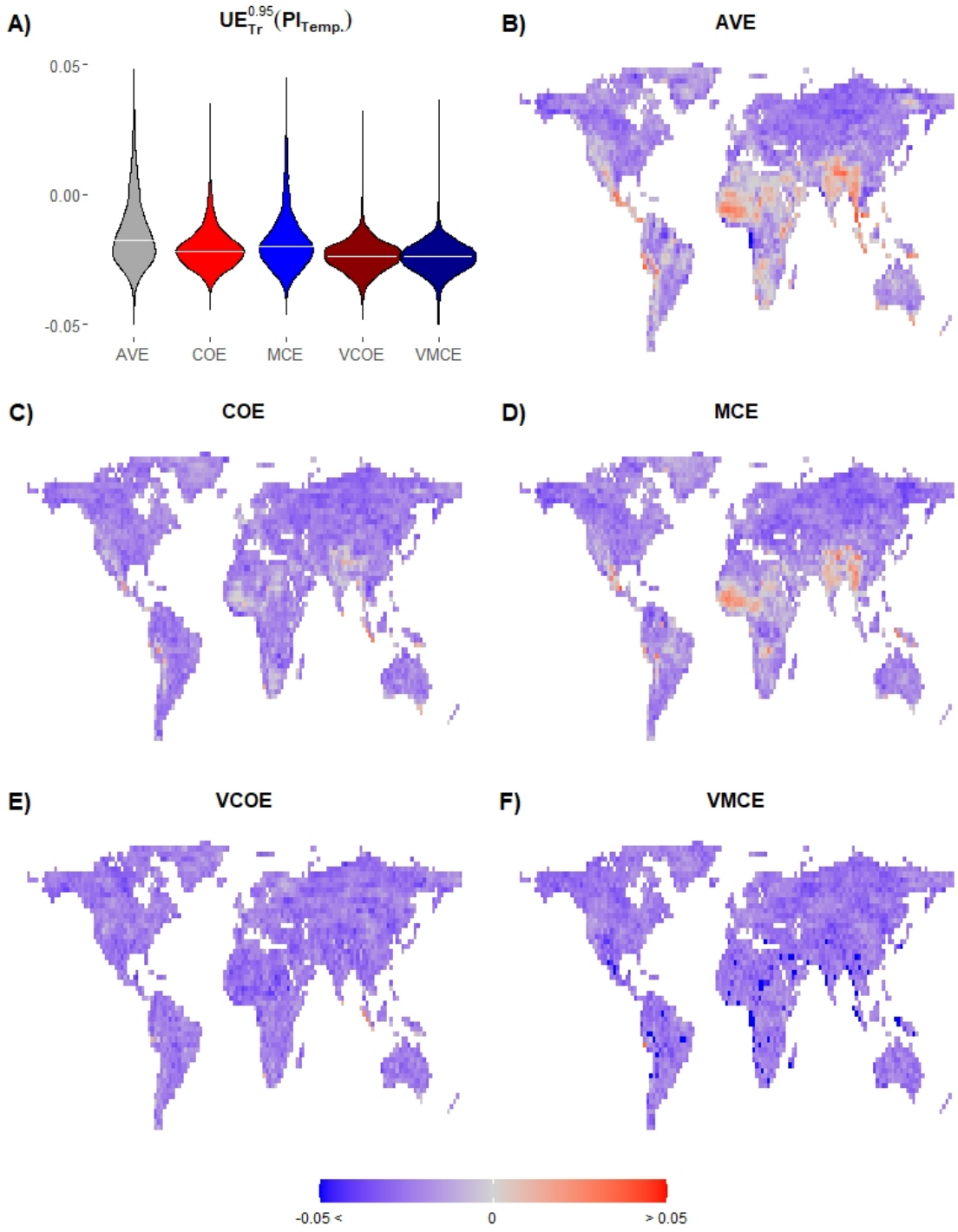

Figure 4.8: Prediction interval uncertainty errors with $1 - \alpha = 0.95$ on training period **(years 1901-1980)** for temperature. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.05 and 0.05. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors respectively.

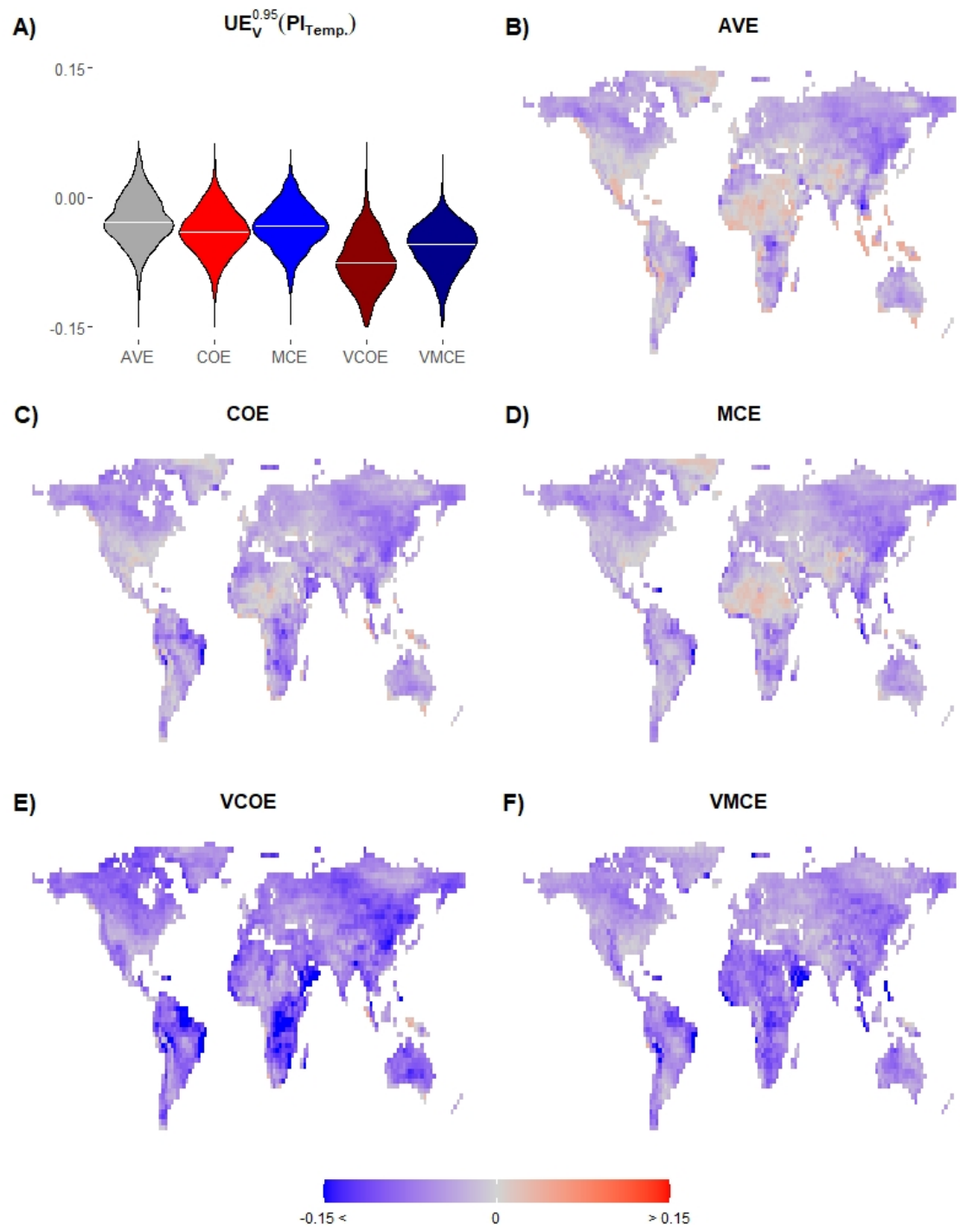

Figure 4.9: Prediction interval uncertainty errors with $1 - \alpha = 0.95$ on validation period **(years 1981-2020)** for temperature. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.15. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors respectively.

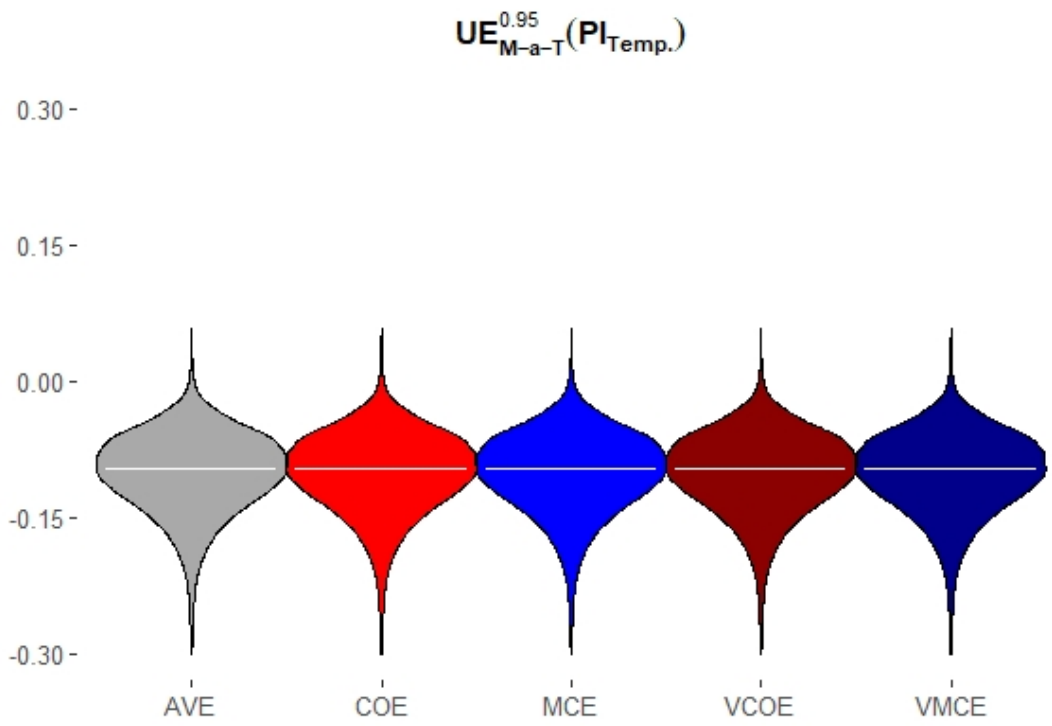

Figure 4.10: Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors with $1-\alpha = 0.95$ in model-as-truth experiments **(years 2021-2100)** for temperature. The Y-axis is cut at -0.30.

The training results show that prediction interval has a very small negative $UE^{0.95}_{Tr}$ across majority of the grid cells for all the ensemble weighting methods. These empirical results show that the proposed method is well-suited for uncertainty estimation of future temperature at $1-\alpha = 0.95$. While the validation results show larger negative $UE^{0.95}_{V}$, especially for VCOE method, the $UE^{0.95}_{V}$ values remain very close to 0 across majority of grid cells confirming the PI method's consistency and reliability. All of the ensemble weighting methods perform at a similar level in model-as-truth experiment at $1-\alpha = 0.95$. As $UE^{0.95}(PI_{Temp.})$ performance is high on validation period but noticeably worse in model-as-truth experiment it indicates that prediction interval is more suitable for near future uncertainty estimation at $1-\alpha = 0.95$. We compare these findings to the weighted quantile interval performance in Tables 4.6, 4.7 and 4.8 as well as Figures 4.11, 4.12, 4.13, 4.14, 4.15,4.16, 4.17, 4.18 and 4.19 below.

4.3.1.2 Weighted quantile interval for temperature

We present the aggregated values of weighted quantile uncertainty error ($UE^{1-\alpha}(WQ_{Temp.})$) on temperature data using all land points weighed according to the grid cells' sizes in Table 4.6 for training period and Table 4.7 for validation period. Those results are accompanied by the weighted quantile interval uncertainty area ($UA^{1-\alpha}(WQ_{Temp.})$) results on temperature data averaged over both training and validation period in Figure 4.8.

$1-\alpha$	*0.65*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*
AVE	-0.12	-0.12	-0.11	-0.11	-0.10	-0.08	**-0.06**
COE	**0.05**	**0.03**	**0.02**	**-0.01**	-0.03	-0.06	-0.09
MCE	-0.07	-0.08	-0.08	-0.08	-0.09	-0.09	-0.10
VCOE	0.10	0.08	0.05	0.02	**-0.01**	**-0.05**	-0.09
VMCE	-0.09	-0.09	-0.10	-0.10	-0.11	-0.12	-0.13

Table 4.6: Average weighted quantile interval uncertainty error results using all land points weighted according to their area sizes on training period **(years 1901-1980)** for temperature. The smallest errors in each column are emphasised in bold.

$1-\alpha$	*0.65*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*
AVE	-0.12	-0.12	-0.11	-0.11	-0.10	-0.08	**-0.05**
COE	**0.03**	**0.02**	**0.00**	**-0.01**	**-0.03**	**-0.06**	-0.09
MCE	-0.08	-0.08	-0.08	-0.08	-0.09	-0.09	-0.09
VCOE	0.05	0.03	0.01	**-0.01**	-0.04	-0.07	-0.10
VMCE	-0.09	-0.10	-0.10	-0.11	-0.11	-0.12	-0.13

Table 4.7: Average weighted quantile interval uncertainty error results using all land points weighted according to their area sizes on validation period **(years 1981-2020)** for temperature. The smallest errors in each column are emphasised in bold.

$1-\alpha$	*0.65*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*
AVE	4.14	4.66	5.24	5.93	6.78	7.97	10.14
COE	4.12	4.55	5.03	5.57	6.20	6.91	7.74
MCE	4.31	4.78	5.32	5.92	6.63	7.57	9.01
VCOE	**3.79**	**4.18**	**4.59**	**5.04**	**5.52**	**6.05**	**6.67**
VMCE	4.29	4.76	5.27	5.85	6.54	7.40	8.48

Table 4.8: Average weighted quantile interval uncertainty area results using all land points weighted according to their area sizes on training and validation period **(years 1901-2020)** for temperature. The smallest area sizes in each column are emphasised in bold.

The results for $1-\alpha \in [0.85, 0.95]$ on training period show that the weighted quantile (WQ) interval is systematically overconfident. The validation period results further confirm that the weighted quantile interval uncertainty error ($UE^{1-\alpha}$) differs from the corresponding $1-\alpha$ values by up to 20% of $1-\alpha$. Furthermore, the range of ($UE^{1-\alpha}$) values in each column is much larger for weighted quantile interval than for prediction interval. This indicates that weighted quantile interval has more limitations than prediction interval on training and validation periods. $UA(WQ)$ is larger for $UA(PI)$ for all $1-\alpha$ values and together with the worse $UE(WQ)$ performance it makes weighted quantile interval less accurate than prediction interval for training and validation periods, especially at high $1-\alpha$.

The detailed weighted quantile ($WQ^{1-\alpha}_{Temp.}$) interval results with $1-\alpha = 0.65, 0.80, 0.95$ are presented as maps of uncertainty error ($UE^{1-\alpha}$) calculated according to Equation 4.5. The results for training (years 1901-1980) and validation (years 1981-2020) periods are followed by model-as-truth experiment summary for validation period (years 2021-2100). Figures 4.11, 4.12 and 4.13 show detailed weighted quantile interval uncertainty error results with $1-\alpha = 0.65$ for temperature below.

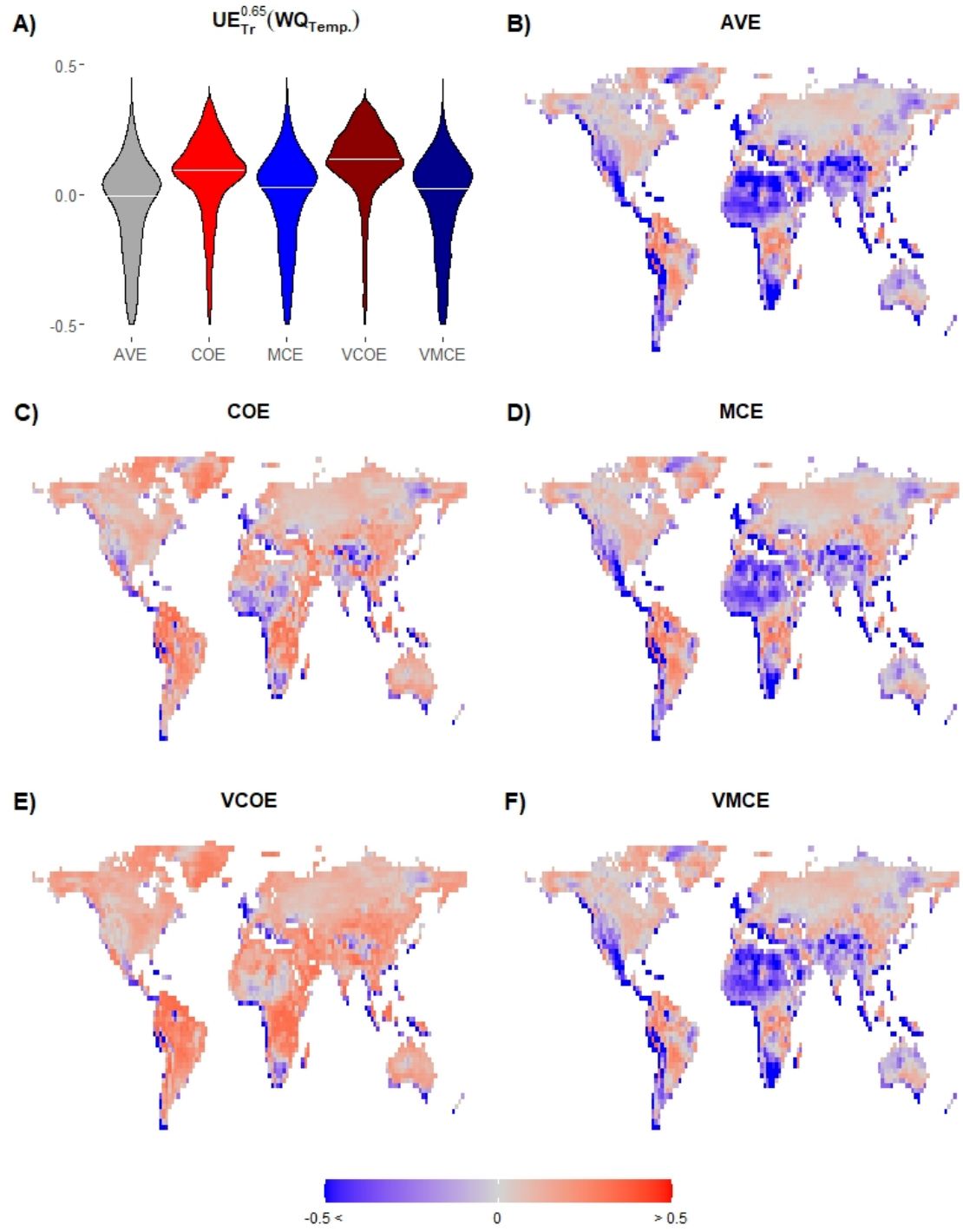

Figure 4.11: Weighted quantile interval uncertainty errors with $1 - \alpha = 0.65$ on training period **(years 1901-1980)** for temperature. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.5. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors respectively.

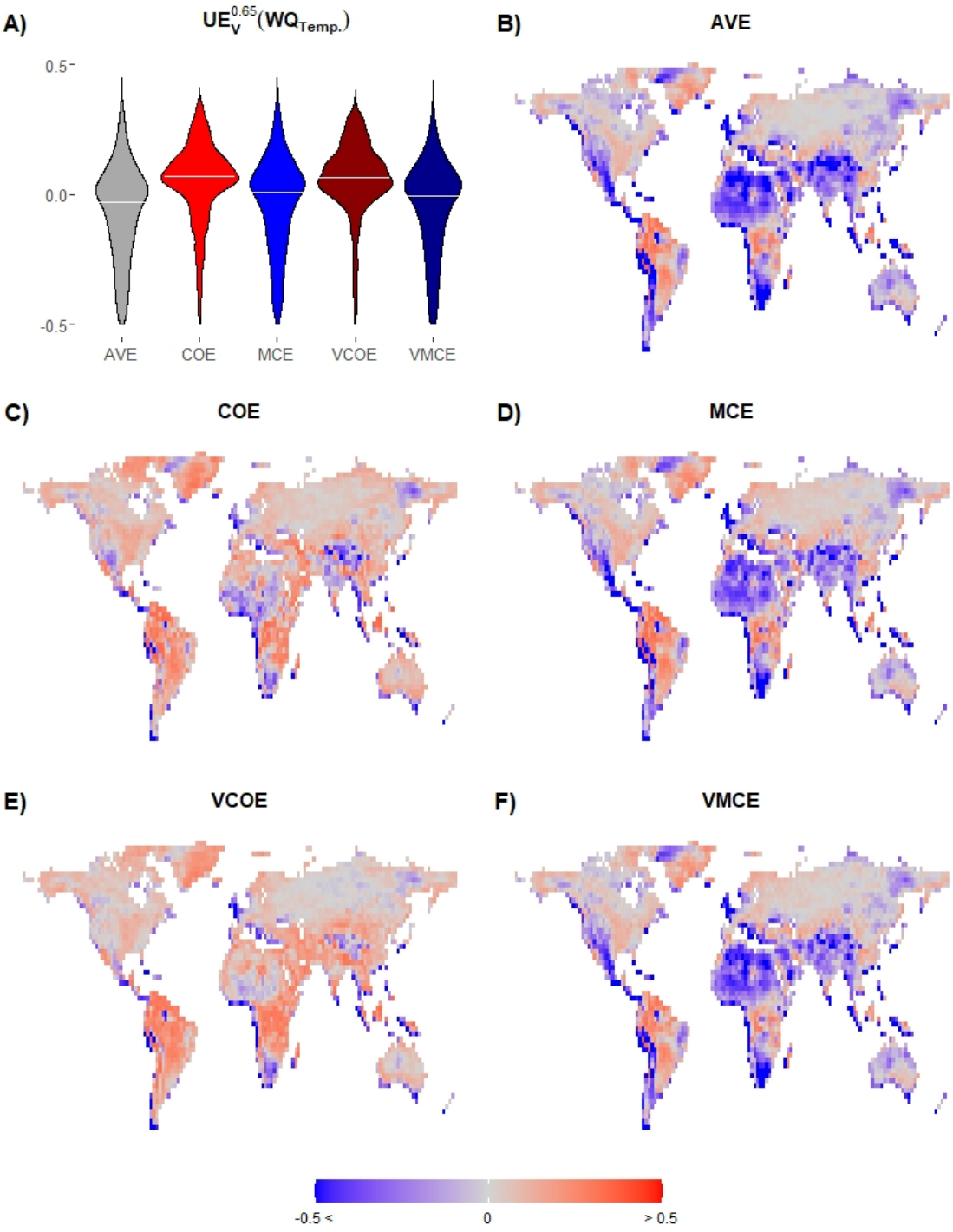

Figure 4.12: Weighted quantile interval uncertainty errors with $1-\alpha = 0.65$ on validation period **(years 1981-2020)** for temperature. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.5. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors respectively.

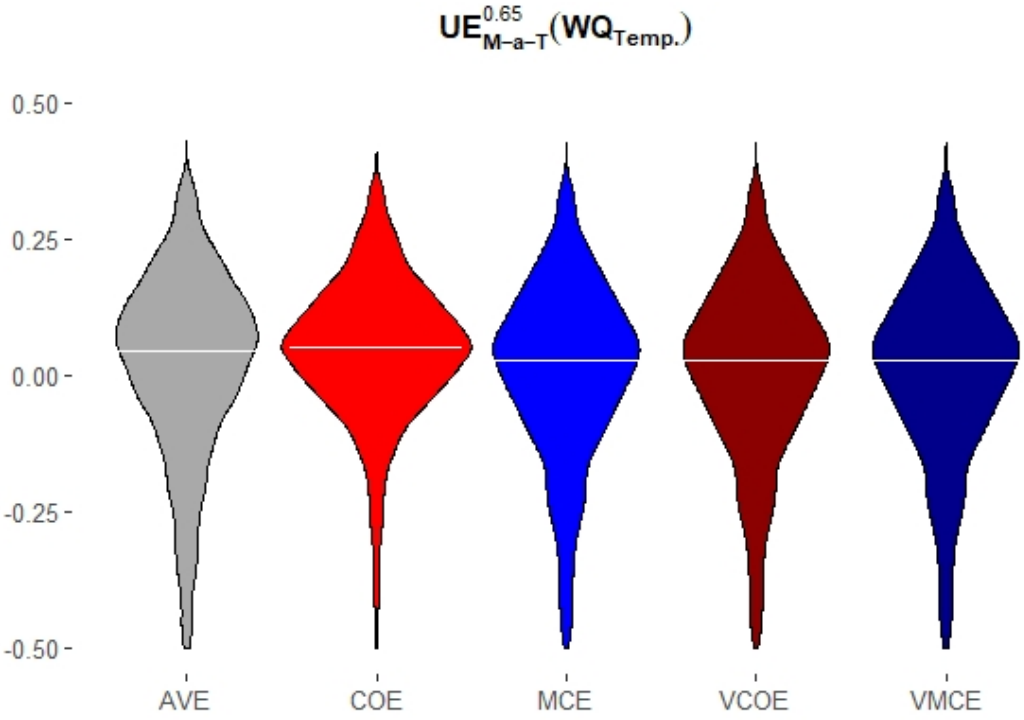

Figure 4.13: Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors with $1-\alpha = 0.65$ in model-as-truth experiments **(years 2021-2100)** for temperature. The Y-axis is cut at -0.5.

Weighted quantile interval has a much larger spread of uncertainty errors on training and validation periods than prediction interval with $1 - \alpha = 0.65$. All ensemble weighting methods perform at similar level on training and validation period as well as in model-as-truth experiment when averaged. As $UE^{0.65}(WQ_{Temp.})$ performance is high in model-as-truth experiment but noticeably worse on validation period it indicates that weighted quantile interval is more suitable for far future uncertainty estimation than the prediction interval at $1 - \alpha = 0.65$. We present uncertainty estimation for $1 - \alpha = 0.8$ in Figures 4.14, 4.15 and 4.16 below.

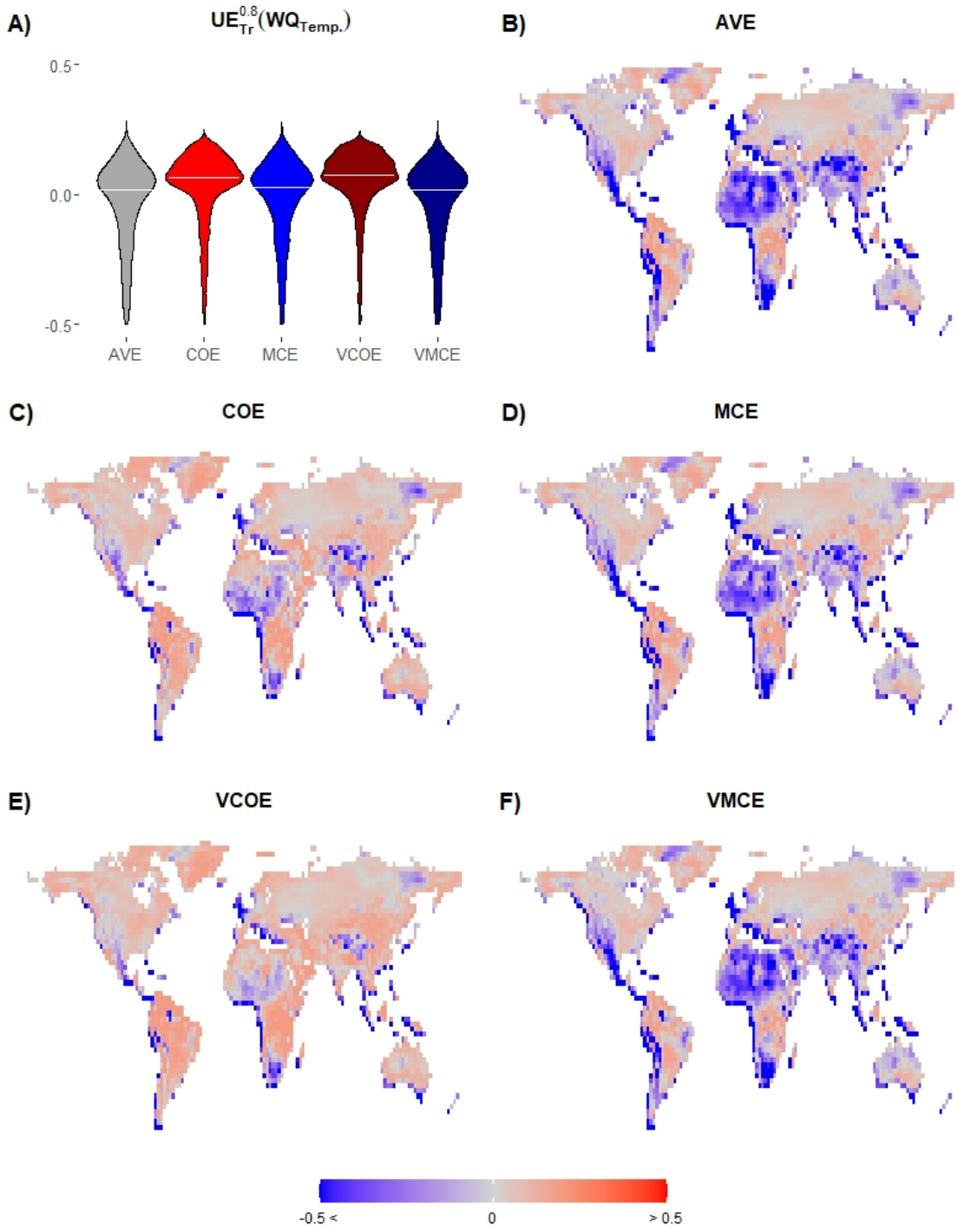

Figure 4.14: Weighted quantile interval uncertainty errors with $1 - \alpha = 0.8$ on training period **(years 1901-1980)** for temperature. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.5. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors respectively.

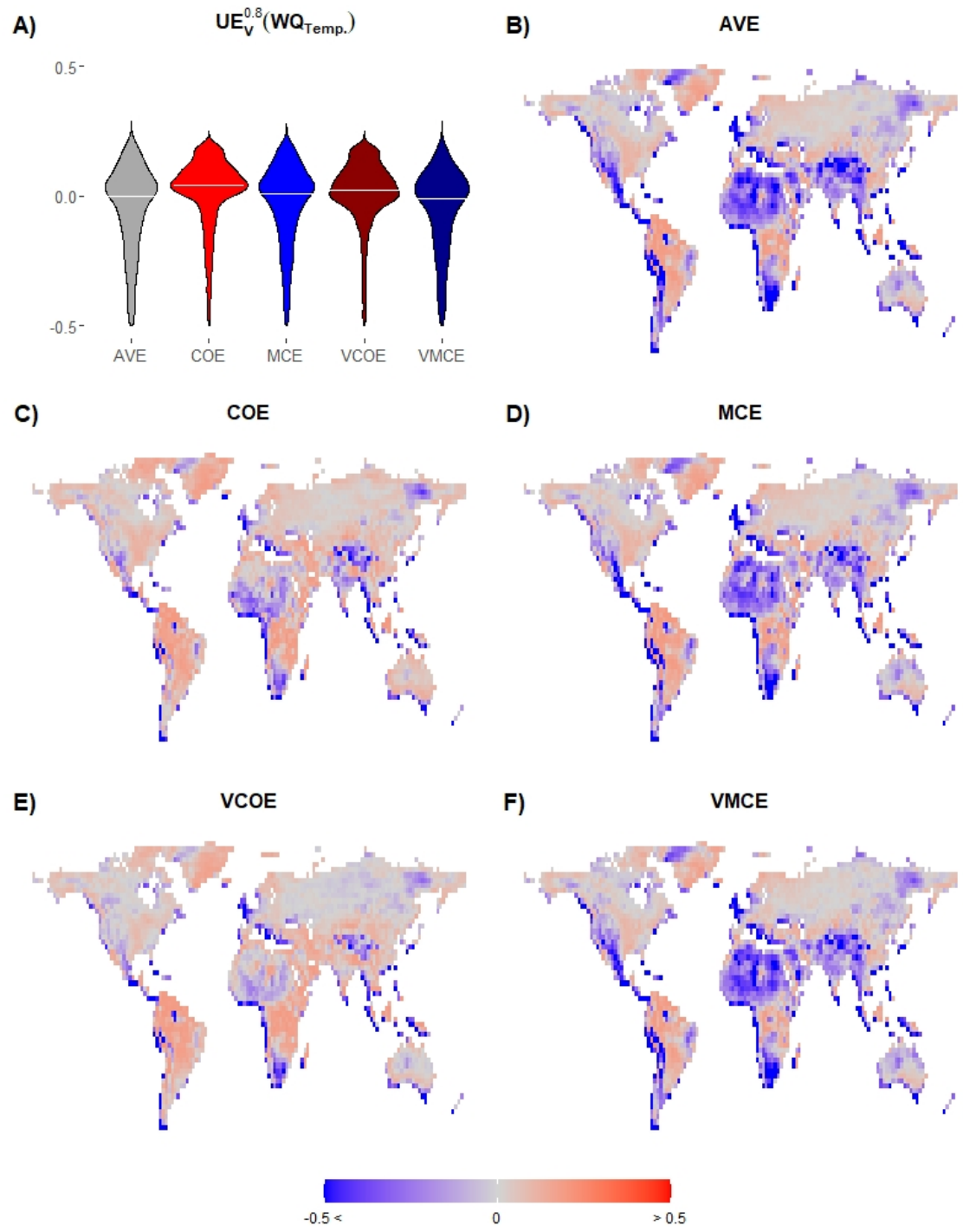

Figure 4.15: Weighted quantile interval uncertainty errors with $1 - \alpha = 0.8$ on validation period **(years 1981-2020)** for temperature. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.5. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors respectively.

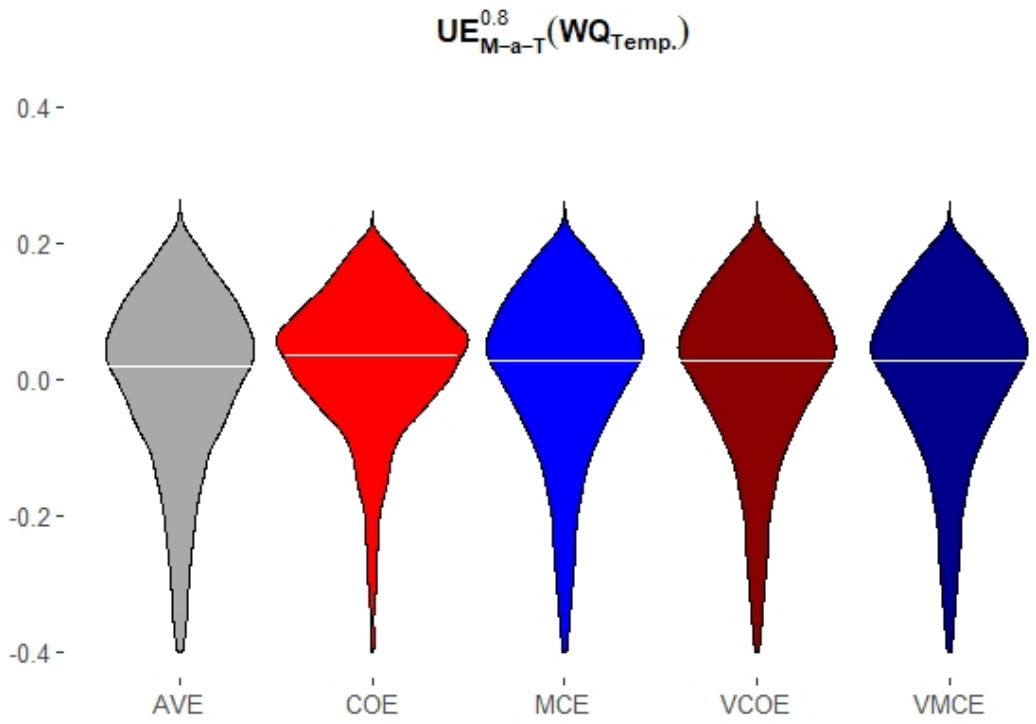

Figure 4.16: Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors with $1-\alpha = 0.8$ in model-as-truth experiments **(years 2021-2100)** for temperature. The Y-axis is cut at -0.4.

The weighted quantile interval results with $1-\alpha = 0.8$ confirm that weighted quantile interval has a much larger spread of uncertainty errors on training and validation periods than prediction interval with $1-\alpha = 0.8$ and that all ensemble weighting methods perform at similar level on training and validation period as well as in model-as-truth experiment when averaged. As $UE^{0.8}(WQ_{Temp.})$ performance is high in model-as-truth as well as on validation period it indicates that weighted quantile interval is more suitable for far future uncertainty estimation than the prediction interval at $1-\alpha = 0.8$.We present $WQ^{0.95}$ results for a commonly used $1-\alpha = 0.95$ in Figures 4.17, 4.18 and 4.19 below.

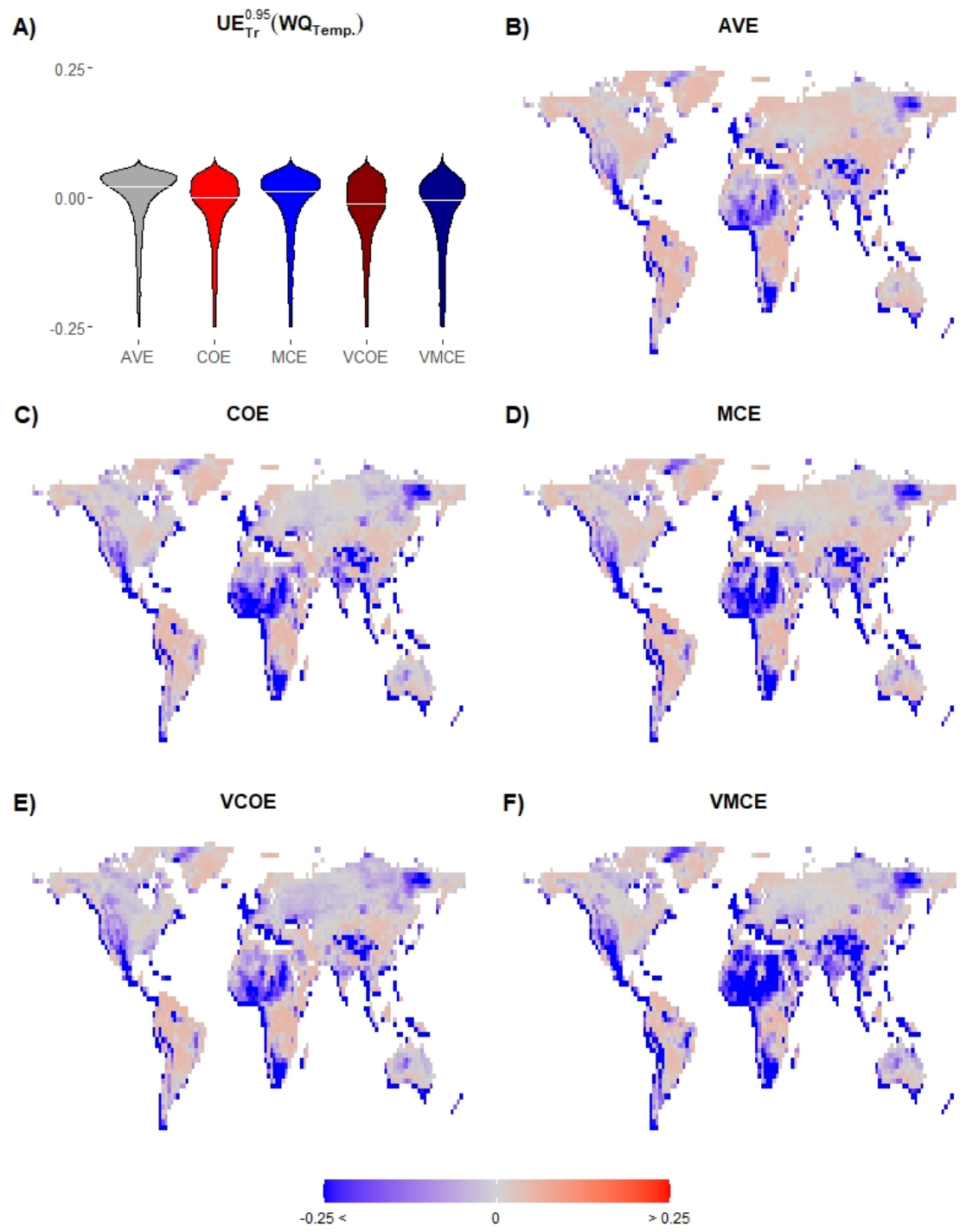

Figure 4.17: Weighted quantile interval uncertainty errors with $1 - \alpha = 0.95$ on training period **(years 1901-1980)** for temperature. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.25. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors respectively.

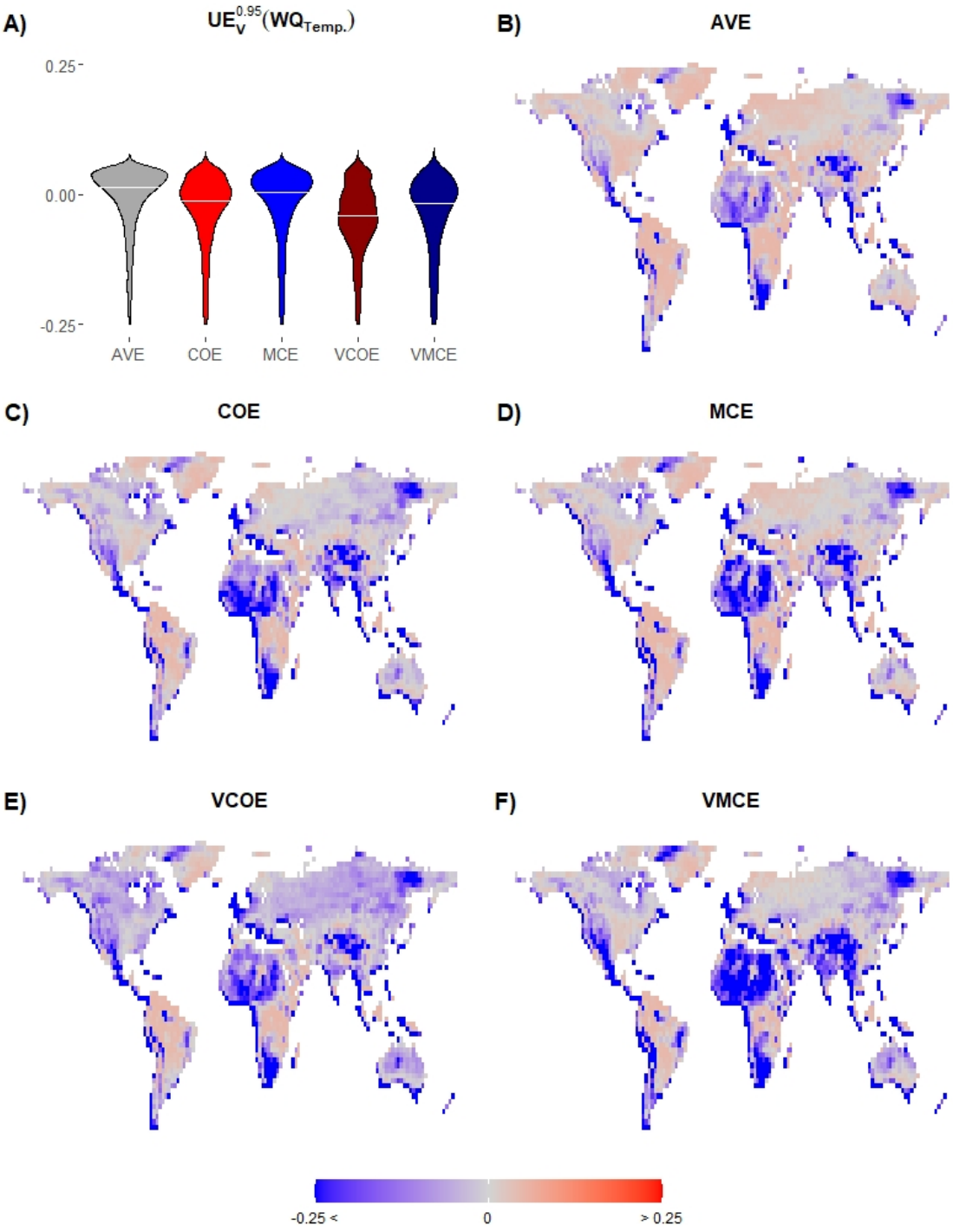

Figure 4.18: Weighted quantile interval uncertainty errors with $1-\alpha = 0.95$ on validation period **(years 1981-2020)** for temperature. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.25. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors respectively.

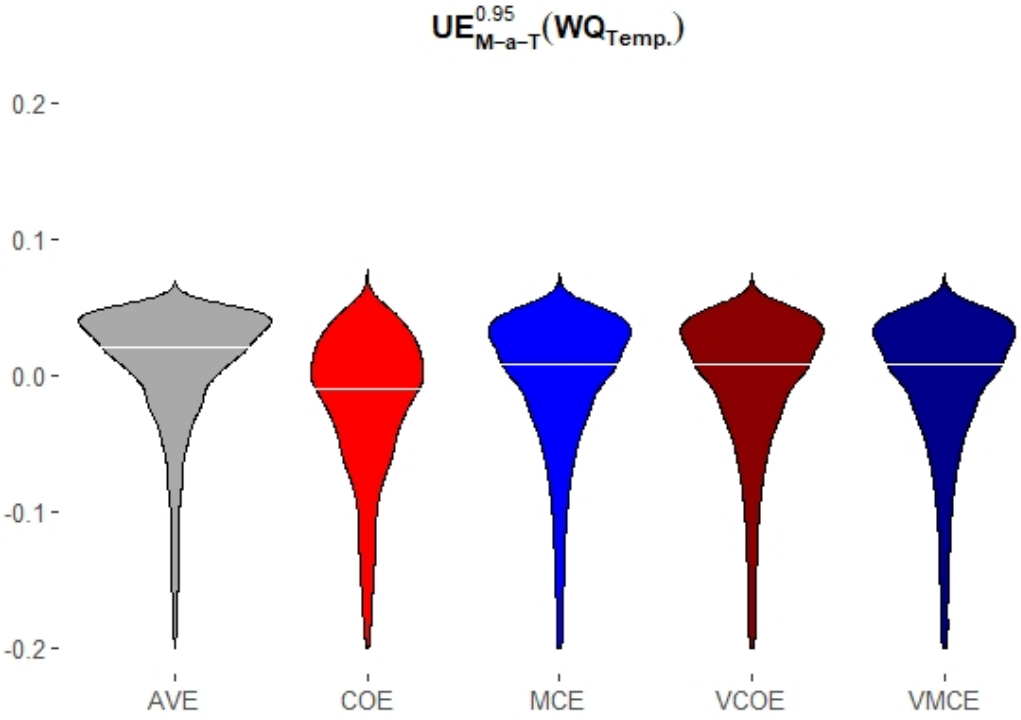

Figure 4.19: Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors with $1-\alpha = 0.95$ in model-as-truth experiments **(years 2021-2100)** for temperature. The Y-axis is cut at -0.20.

The WQ uncertainty error $UE^{0.95}$ results are nearly the same for all the methods with VCOE having slightly more negative UE values on training and validation periods. The geographical patterns for positive and negative UE values are very similar for all the methods on training and validation periods. All ensemble weighting methods perform at similar level in model-as-truth experiment when averaged. As $UE^{0.95}(WQ_{Temp.})$ performance is high in model-as-truth experiment but noticeably worse on validation period it indicates that weighted quantile interval is more suitable for far future uncertainty estimation than the prediction interval at $1-\alpha = 0.95$.

4.3.2 Precipitation data

4.3.2.1 Prediction interval for precipitation data

We present the aggregated values of prediction interval uncertainty error ($UE^{1-\alpha}(PI_{Precip.})$) on precipitation data using all land points weighed according to the grid cells' sizes in Table 4.9 for training period and Table 4.10 for validation period. Those results are accompanied by the prediction interval uncertainty area ($UA^{1-\alpha}(PI_{Precip.})$) results on precipitation data averaged over both training and validation period in Table 4.11.

$1-\alpha$	*0.65*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*
AVE	**0.07**	**0.05**	**0.04**	**0.02**	**0.01**	-0.01	**-0.02**
COE	0.09	0.07	0.05	0.03	**0.01**	-0.01	**-0.02**
MCE	0.08	0.06	0.05	0.03	**0.01**	-0.01	**-0.02**
VCOE	0.10	0.08	0.06	0.04	0.02	**0.00**	**-0.02**
VMCE	0.10	0.09	0.06	0.04	0.02	**0.00**	**-0.02**

Table 4.9: Average prediction interval uncertainty error results using all land points weighted according to their area sizes on training period **(years 1901-1980)** for precipitation. The smallest errors in each column are emphasised in bold.

$1-\alpha$	*0.65*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*
AVE	0.04	**0.02**	0.01	**-0.01**	**-0.02**	**-0.04**	**-0.04**
COE	0.04	0.03	0.01	**-0.01**	-0.03	**-0.04**	-0.05
MCE	0.04	**0.02**	0.01	**-0.01**	-0.03	**-0.04**	-0.05
VCOE	0.04	**0.02**	**0.00**	**-0.01**	-0.03	-0.05	-0.06
VMCE	**0.03**	**0.02**	**0.00**	-0.02	-0.04	-0.05	-0.06

Table 4.10: Average prediction interval uncertainty error results using all land points weighted according to their area sizes on validation period **(years 1981-2020)** for precipitation. The smallest errors in each column are emphasised in bold.

$1-\alpha$	*0.65*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*
AVE	73.18	79.97	87.31	95.49	105.04	117.10	135.05
COE	66.66	73.09	80.03	87.77	96.80	108.19	125.13
MCE	67.14	73.57	80.51	88.26	97.30	108.70	125.64
VCOE	61.00	67.29	74.08	81.66	90.49	101.64	118.20
VMCE	**55.61**	**61.41**	**67.67**	**74.65**	**82.80**	**93.07**	**108.34**

Table 4.11: Average prediction interval uncertainty area results using all land points weighted according to their area sizes on training and validation period **(years 1901-2020)** for precipitation. The smallest area sizes in each column are emphasised in bold.

The results for the training period show that the prediction interval reflects $1-\alpha$ values method adequately for all the methods in this study with AVE having a slight advantage. This is confirmed by the results on validation period. All of the methods are relatively close in $UE^{1-\alpha}$ metric, but have a significant difference in $UA^{1-\alpha}$ metric with VMCE method having the most narrow prediction interval. This confirms the previous findings that VMCE has a relatively high performance in terms of overall results across both $UE^{1-\alpha}$ and $UA^{1-\alpha}$ metrics.

The detailed prediction interval ($PI^{1-\alpha}_{Precip.}$) results with $1-\alpha = 0.65, 0.80, 0.95$ are presented as maps of uncertainty error ($UE^{1-\alpha}$) calculated according to Equation 4.5. The results for training (years 1901-1980) and validation (years 1981-2020) periods are followed by model-as-truth experiment summary for validation period (years 2021-2100). Figures 4.20,4.21 and 4.22 show detailed prediction interval uncertainty error results with $1-\alpha = 0.65$ for precipitation below.

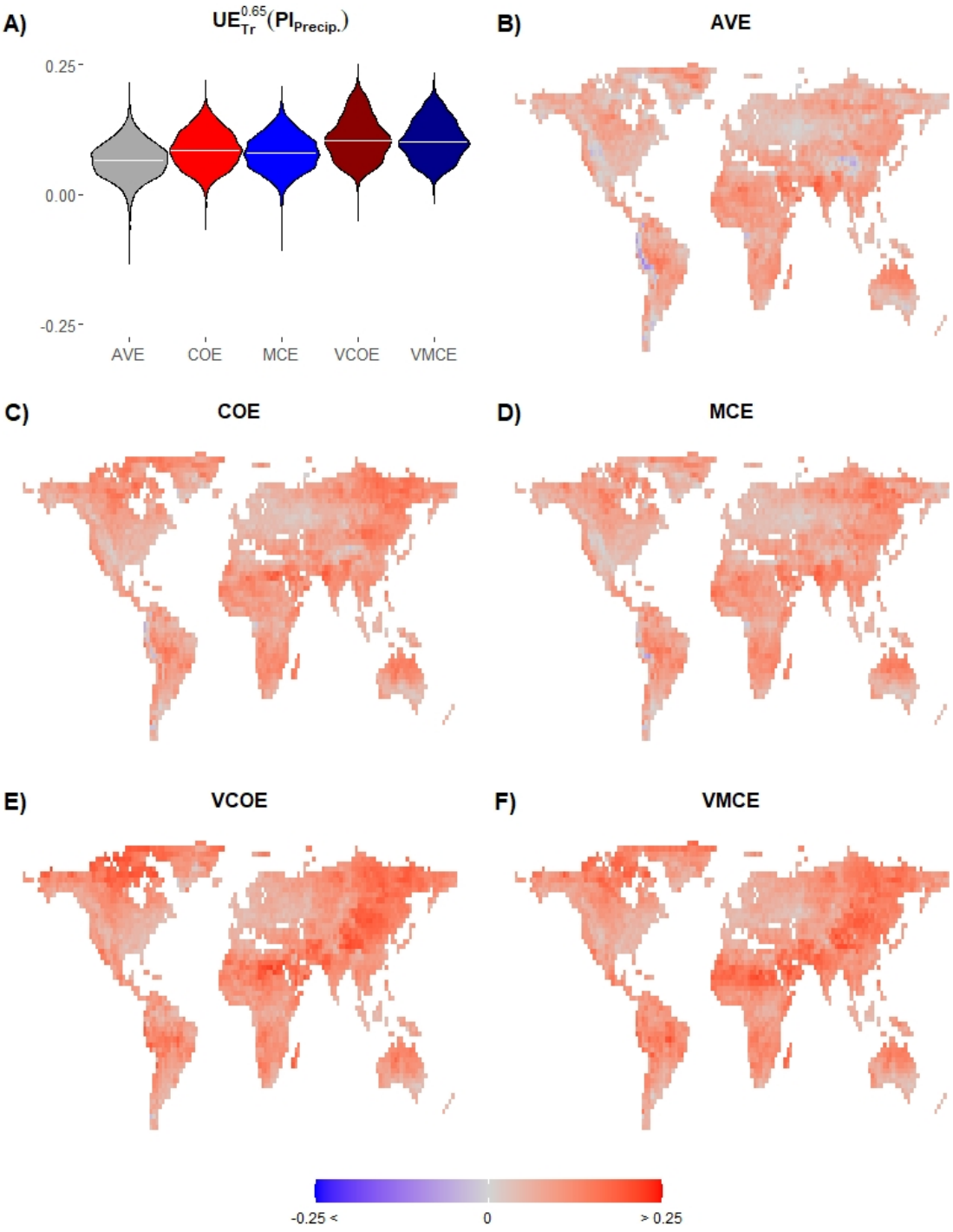

Figure 4.20: Prediction interval uncertainty errors with $1 - \alpha = 0.65$ on training period **(years 1901-1980)** for precipitation. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at 0.25. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors respectively.

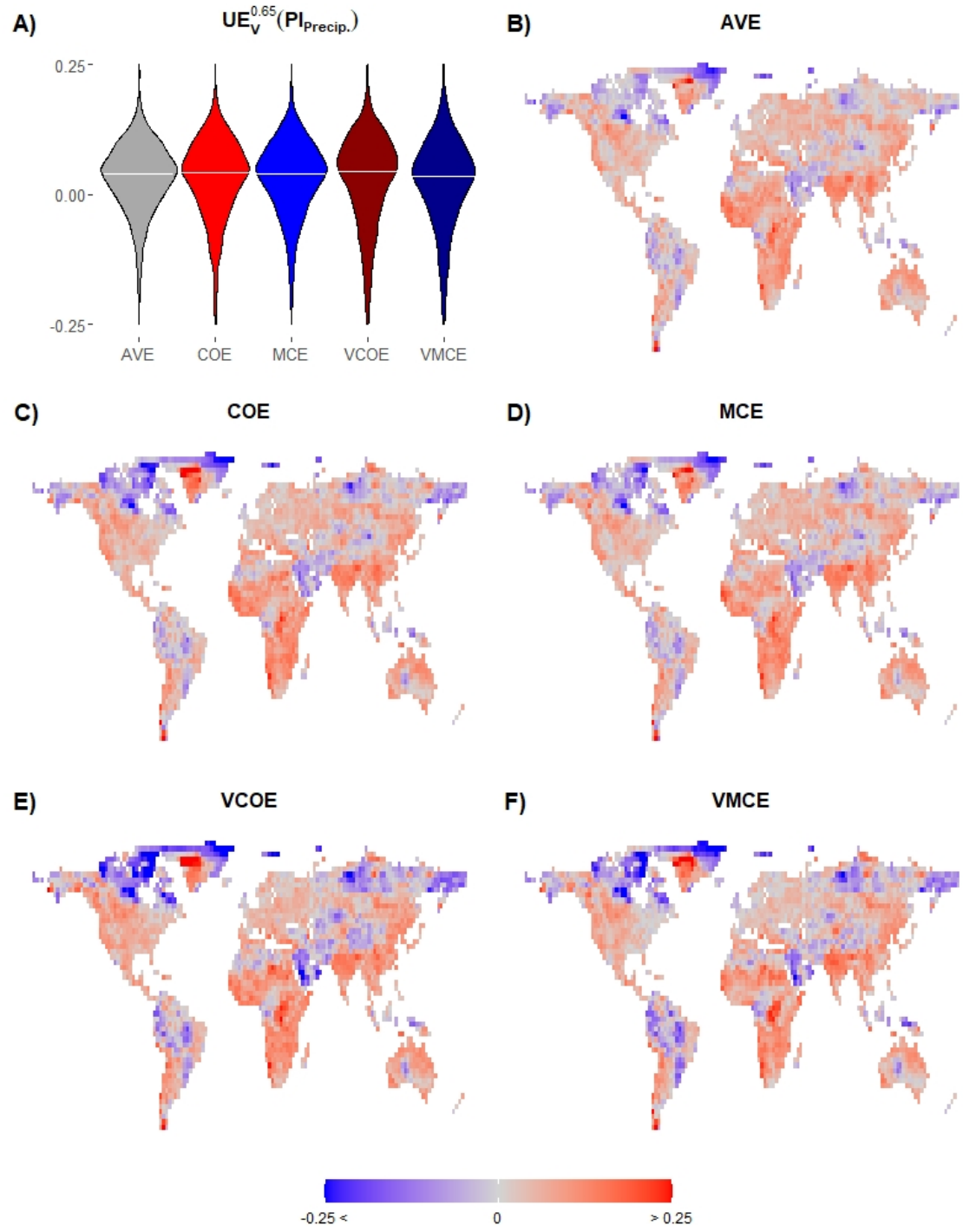

Figure 4.21: Prediction interval uncertainty errors with $1 - \alpha = 0.65$ on validation period **(years 1981-2020)** for precipitation. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.25 and 0.25. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors respectively.

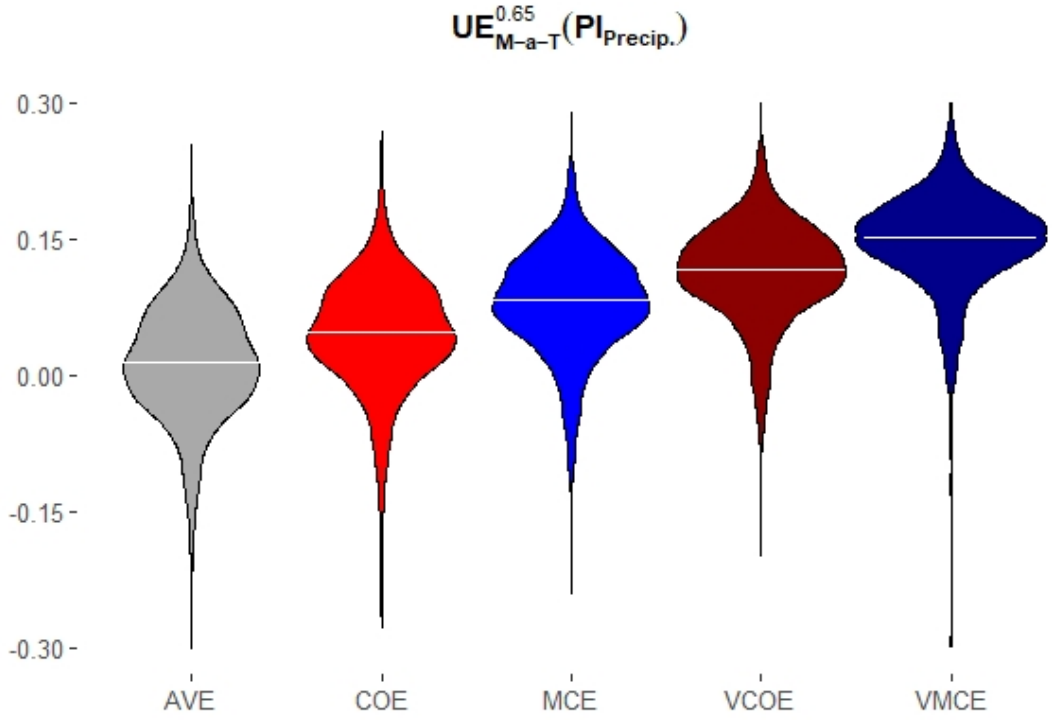

Figure 4.22: Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors with $1 - \alpha = 0.65$ in model-as-truth experiments **(years 2021-2100)** for precipitation. The Y-axis is cut at -0.30 and 0.30.

The geo-spatial distribution shows a slight advantage of the AVE method with all methods consistently having positive $UE^{0.65}_{Tr}$ values across the majority of grid cells on training data. All ensemble weighted methods have similar performance on validation period with similar regional distribution of positive and negative $UE^{0.65}_{V}$ values. The model-as-truth experiment results for precipitation confirm the results for temperature (See Figure 4.4). The spread of $UE^{0.65}_{M-a-T}$ for precipitation is significantly smaller than the spread of $UE^{0.65}_{M-a-T}$ for temperature. As $UE^{0.65}(PI_{Precip.})$ performance is high on validation period but noticeably worse in model-as-truth experiment it indicates that prediction interval is more suitable for near future uncertainty estimation at $1 - \alpha = 0.65$.

We compare these $PI^{0.65}$ results results with $PI^{0.8}$ results for $1 - \alpha = 0.8$ in Figures 4.23, 4.24 and 4.25 below.

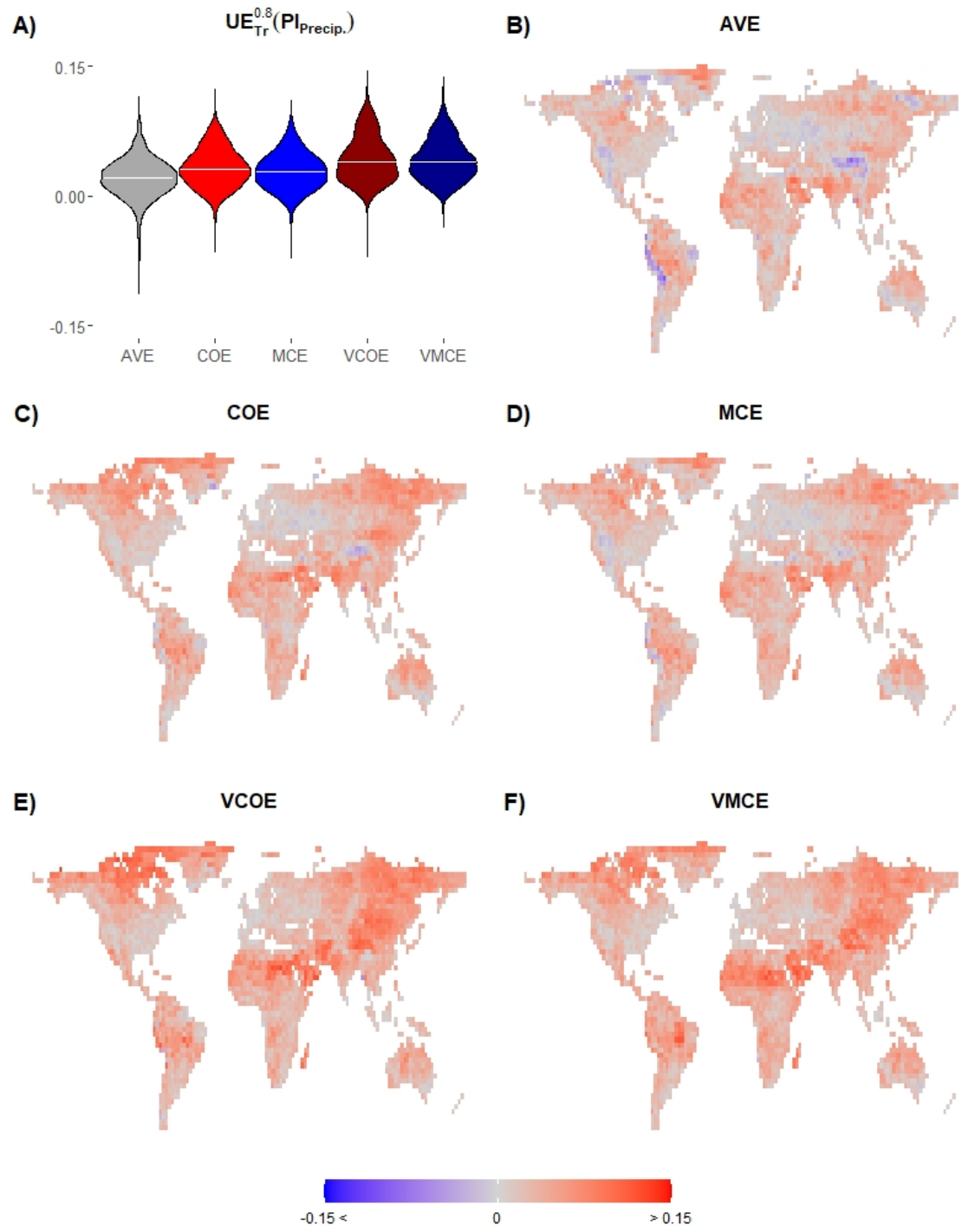

Figure 4.23: Prediction interval uncertainty errors with $1 - \alpha = 0.8$ on training period **(years 1901-1980)** for precipitation. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at 0.15. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors respectively.

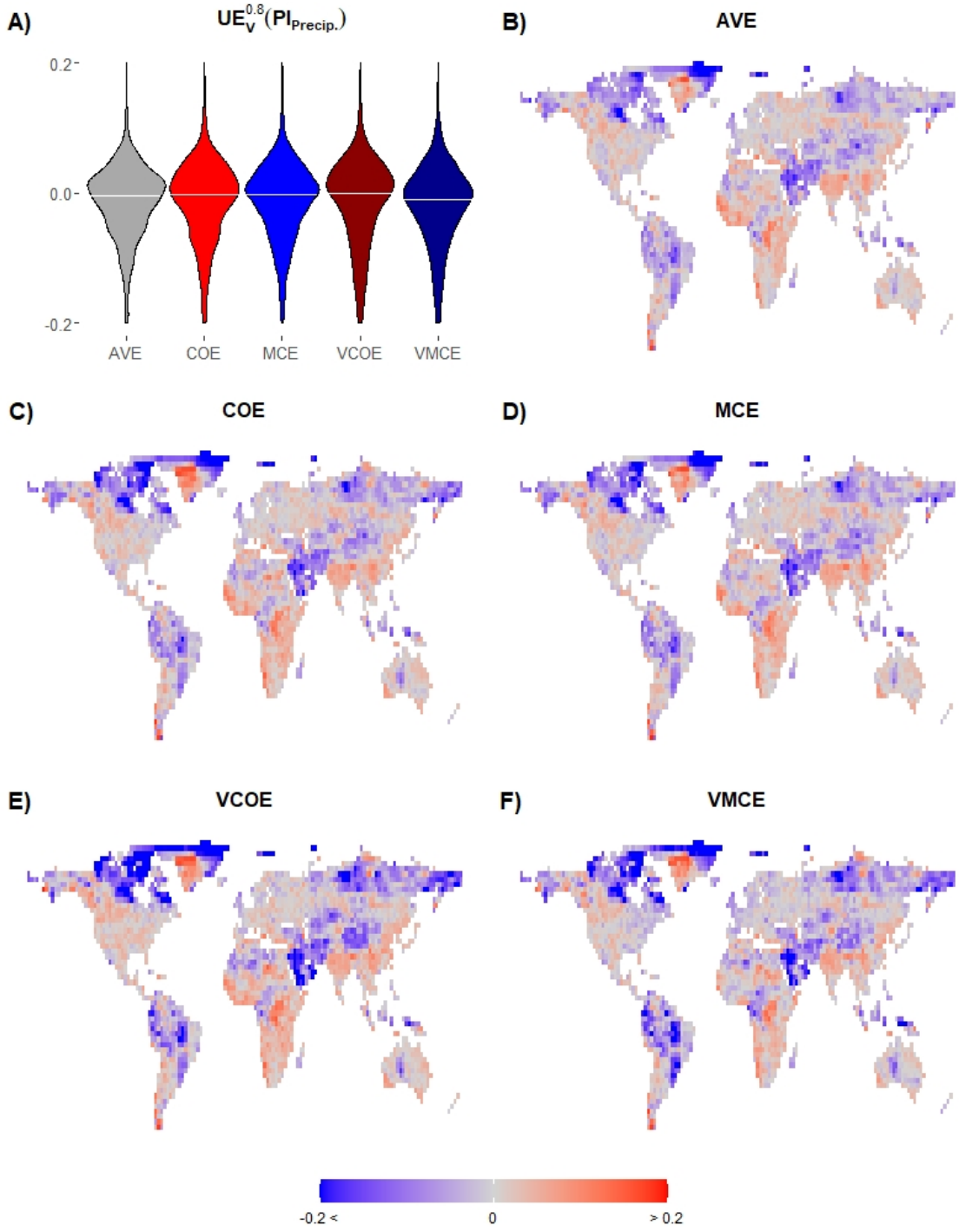

Figure 4.24: Prediction interval uncertainty errors with $1 - \alpha = 0.8$ on validation period **(years 1981-2020)** for precipitation. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.2 and 0.2. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors respectively.

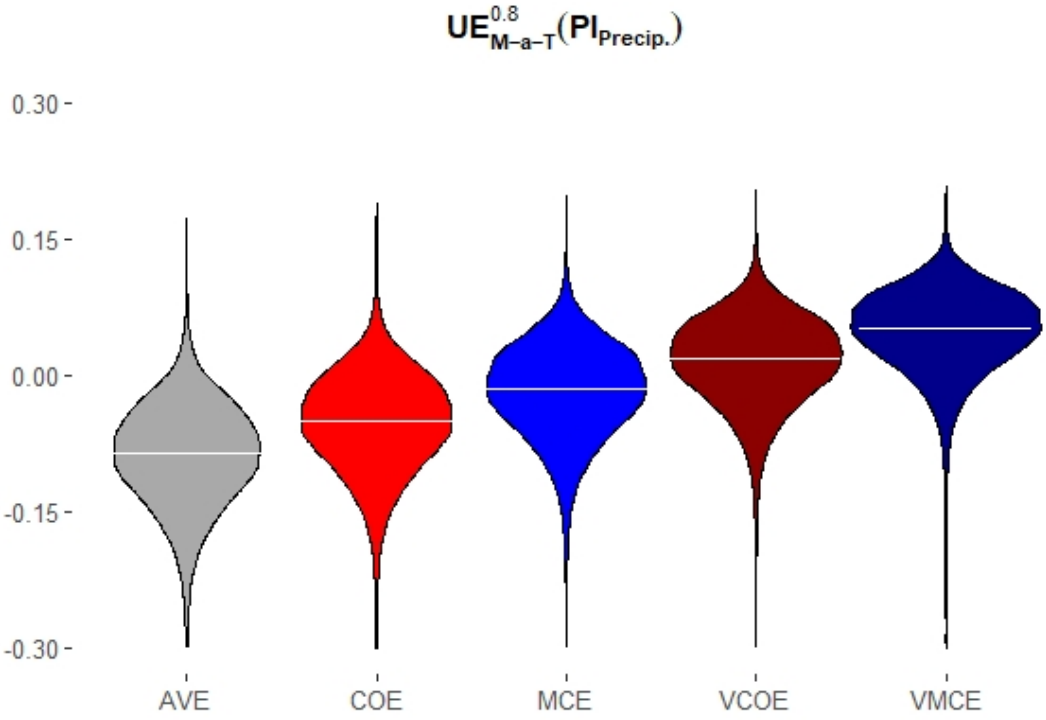

Figure 4.25: Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors with $1-\alpha = 0.8$ in model-as-truth experiments **(years 2021-2100)** for precipitation. The Y-axis is cut at -0.30.

The training data shows that $UE^{0.8}_{Tr}$ results are consistently closer to $1-\alpha = 0.8$ across the globe compared to the results with $1-\alpha = 0.65$ on training data. The validation data and model-as-truth experiment results confirm the findings for $1-\alpha = 0.65$, with MCE having the best performance in model-as-truth experiment. As $UE^{0.8}(PI_{Precip.})$ performance is high on validation period but noticeably worse in model-as-truth experiment it indicates that prediction interval is more suitable for near future uncertainty estimation at $1-\alpha = 0.8$.

We present $PI^{0.95}$ results for a commonly used $1-\alpha = 0.95$ in Figures 4.26, 4.27 and 4.28 below.

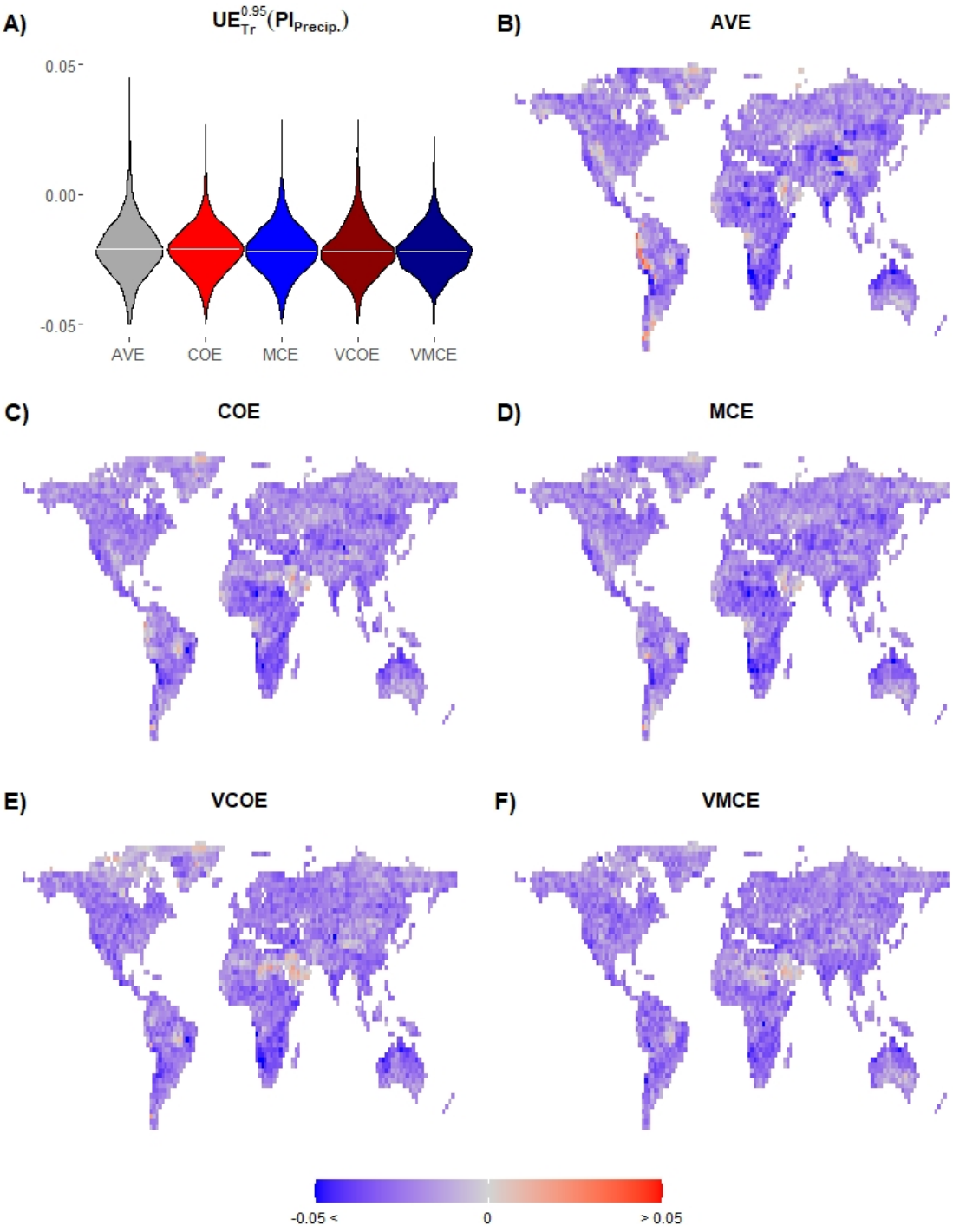

Figure 4.26: Prediction interval uncertainty errors with $1 - \alpha = 0.95$ on training period **(years 1901-1980)** for precipitation. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.05. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors respectively.

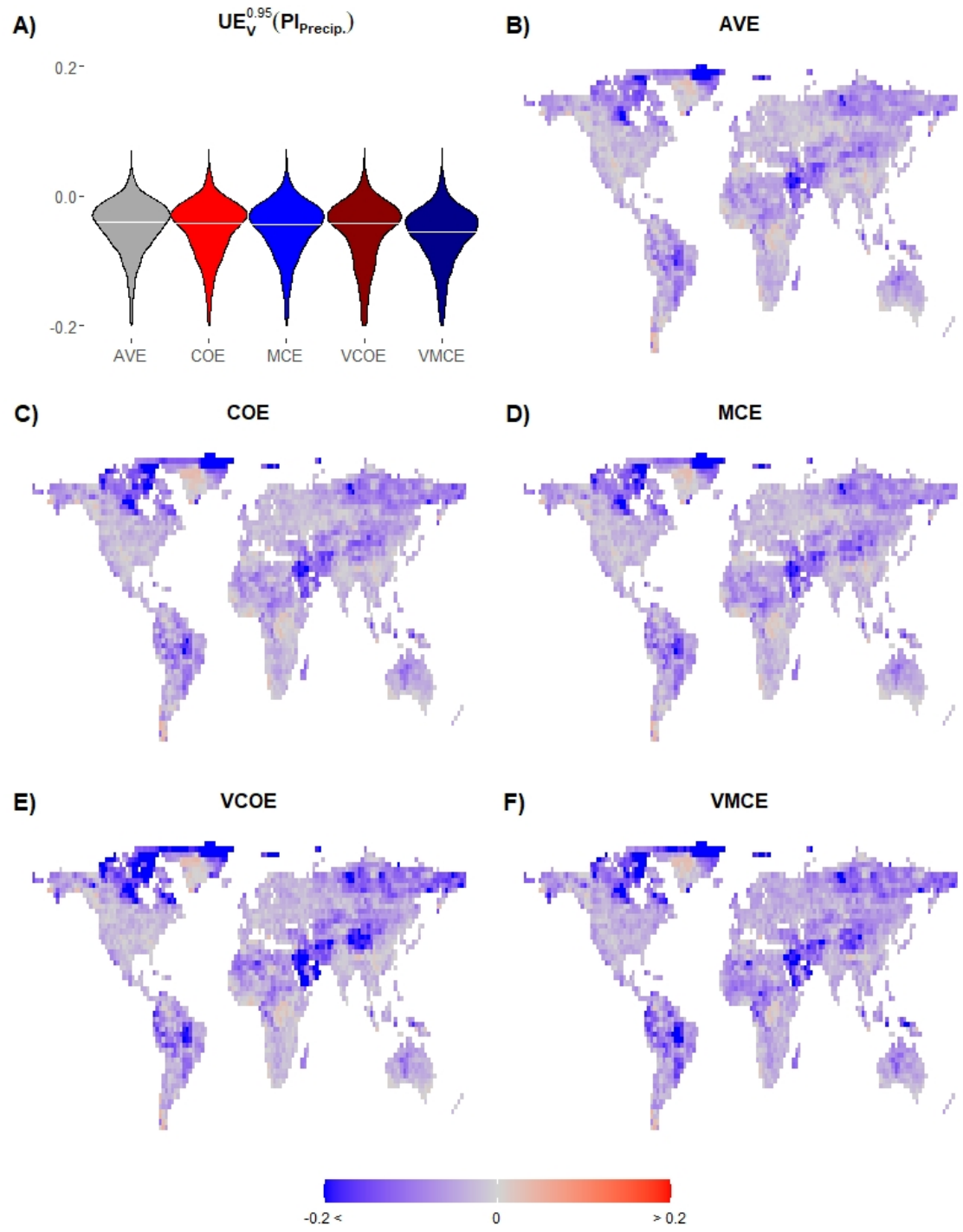

Figure 4.27: Prediction interval uncertainty errors with $1 - \alpha = 0.95$ on validation period **(years 1981-2020)** for precipitation. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.20. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors respectively.

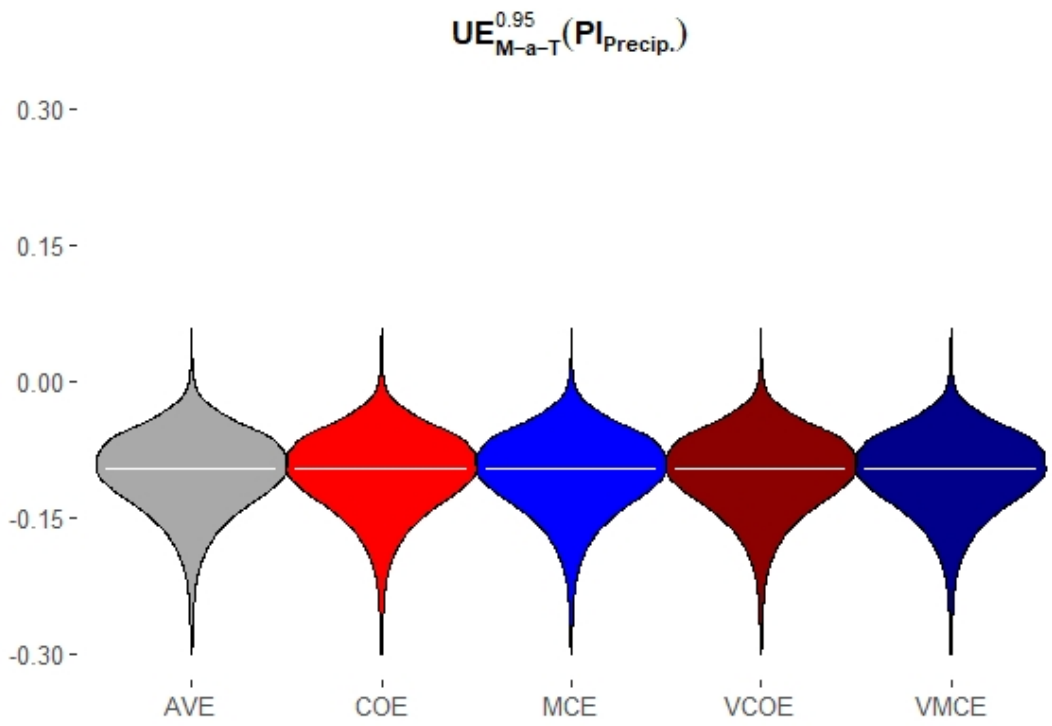

Figure 4.28: Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' prediction interval uncertainty errors with $1 - \alpha = 0.95$ in model-as-truth experiments **(years 2021-2100)** for precipitation. The Y-axis is cut at -0.30.

The training results show that prediction interval has a very small negative $UE^{0.95}_{Tr}$ across majority of the grid cells for all the ensemble weighting methods. These empirical results show that the proposed method is accurate for precipitation uncertainty estimation on training period at $1 - \alpha = 0.95$. While the validation results show larger negative $UE^{0.95}_{V}$, especially for VMCE method, the $UE^{0.95}_{V}$ values remain very close to 0 across majority of grid cells confirming the PI method's consistency and reliability. All of the ensemble weighting methods perform at a similar level in model-as-truth experiment confirming the $UE^{0.95}_{M-a-T}$ results for temperature. As $UE^{0.95}(PI_{Precip.})$ performance is high on validation period but noticeably worse in model-as-truth experiment it indicates that prediction interval is more suitable for near future uncertainty estimation at $1 - \alpha = 0.95$. We compare these findings to the weighted quantile interval performance in Tables 4.12, 4.13 and 4.14 as well as Figures 4.29, 4.30, 4.31, 4.32, 4.33, 4.34, 4.35, 4.36 and 4.37 below.

4.3.2.2 Weighted quantile interval for precipitation

We present the aggregated values of weighted quantile interval uncertainty error ($UE^{1-\alpha}(WQ_{Precip.})$) on precipitation data using all land points weighed according to the grid cells' sizes in Table 4.12 for training period and Table 4.13 for validation period. Those results are accompanied by the weighted quantile interval uncertainty area ($UA^{1-\alpha}(WQ_{Precip.})$) results on precipitation data averaged over both training and validation period in Figure 4.14.

$1-\alpha$	*0.65*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*
AVE	**-0.01**	**-0.01**	**0.00**	**0.00**	**0.01**	0.02	0.03
COE	0.14	0.12	0.11	0.09	0.07	0.05	0.02
MCE	0.02	0.02	0.02	0.01	**0.01**	0.01	**0.01**
VCOE	0.12	0.11	0.10	0.08	0.06	0.04	**0.01**
VMCE	0.02	0.02	0.02	0.02	**0.01**	**0.00**	**-0.01**

Table 4.12: Average weighted quantile interval interval uncertainty error results using all land points weighted according to their area sizes on training period **(years 1901-1980)** for precipitation. The smallest errors in each column are emphasised in bold.

$1-\alpha$	*0.65*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*
AVE	-0.02	-0.02	-0.01	-0.01	**0.00**	**0.01**	0.03
COE	0.13	0.11	0.10	0.08	0.06	0.04	0.02
MCE	**0.01**	0.01	**0.00**	**0.00**	**0.00**	**0.01**	**0.01**
VCOE	0.10	0.09	0.08	0.06	0.05	0.03	0.01
VMCE	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**-0.01**	**-0.01**

Table 4.13: Average weighted quantile interval uncertainty error results using all land points weighted according to their area sizes on validation period **(years 1981-2020)** for precipitation. The smallest errors in each column are emphasised in bold.

$1-\alpha$	*0.65*	*0.7*	*0.75*	*0.8*	*0.85*	*0.9*	*0.95*
AVE	**68.32**	**77.33**	**87.66**	100.03	115.82	138.78	187.18
COE	75.88	84.11	93.61	104.52	117.56	134.64	160.20
MCE	71.33	79.49	88.81	99.84	113.50	132.47	168.98
VCOE	76.26	84.24	93.09	103.27	115.59	131.74	156.15
VMCE	70.77	78.91	88.19	**99.09**	**112.60**	**131.10**	**156.21**

Table 4.14: Average weighted quantile interval uncertainty area results using all land points weighted according to their area sizes on training and validation period **(years 1901-2020)** for precipitation. The smallest area sizes in each column are emphasised in bold.

The results for $1-\alpha \in [0.65, 0.75]$ on training period show that the weighted quantile (WQ) interval is too conservative for COE and VCOE methods. The validation period results further confirm that the weighted quantile interval uncertainty error ($UE^{1-\alpha}$) differs from the corresponding $1-\alpha$ values for up to 20% of $1-\alpha$. Furthermore, the range of ($UE^{1-\alpha}$) values in each column is much larger for weighted quantile interval than for prediction interval. This confirms the previous findings that weighted quantile interval is less accurate than prediction interval on training data. $UA(WQ)$ is larger for $UA(PI)$ for all $1-\alpha$ values and together with the worse $UE(WQ)$ performance it makes weighted quantile interval less suitable than prediction interval for future precipitation uncertainty estimation, especially at low $1-\alpha$.

The detailed weighted quantile ($WQ^{1-\alpha}_{Precip.}$) interval results with $1-\alpha = 0.65, 0.80, 0.95$ are presented as maps of uncertainty error ($UE^{1-\alpha}$) calculated according to Equation 4.5. The results for training (years 1901-1980) and validation (years 1981-2020) periods are followed by model-as-truth experiment summary for validation period (years 2021-2100). Figures 4.29, 4.30 and 4.31 show detailed weighted quantile interval uncertainty error results with $1-\alpha = 0.65$ for precipitation below.

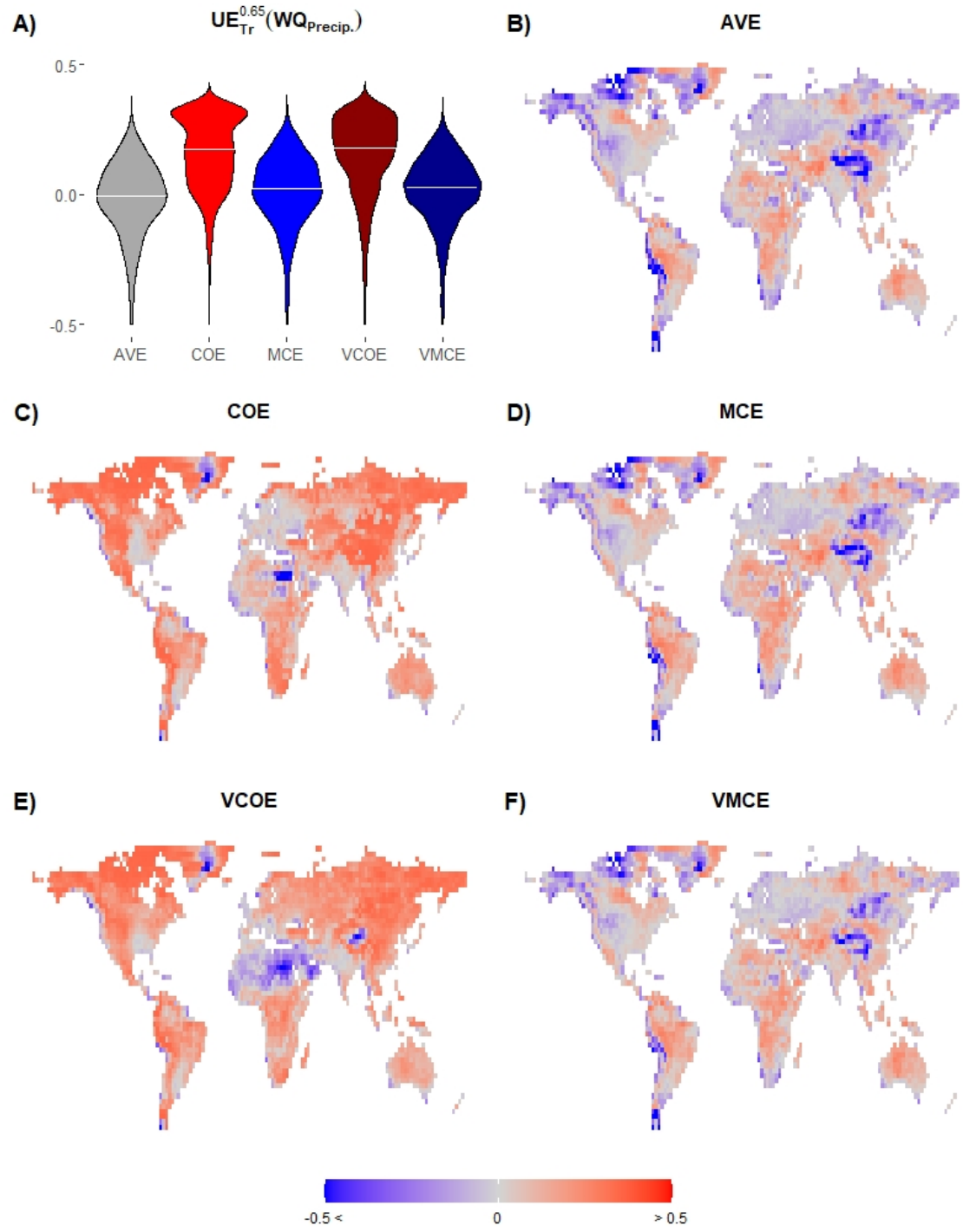

Figure 4.29: Weighted quantile interval uncertainty errors with $1 - \alpha = 0.65$ on training period **(years 1901-1980)** for precipitation. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.5. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors respectively.

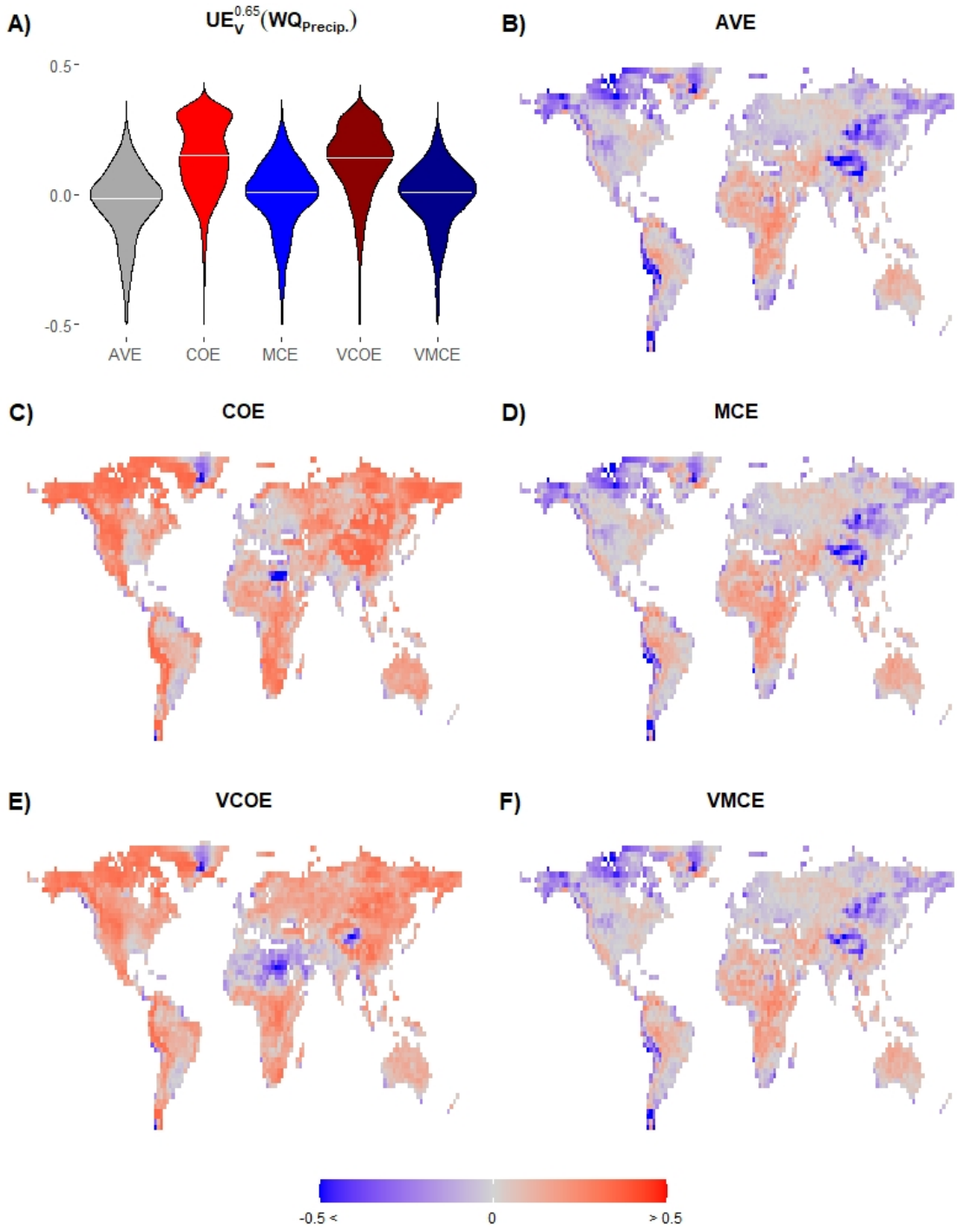

Figure 4.30: Weighted quantile interval uncertainty errors with $1-\alpha = 0.65$ on validation period **(years 1981-2020)** for precipitation. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.5. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors respectively.

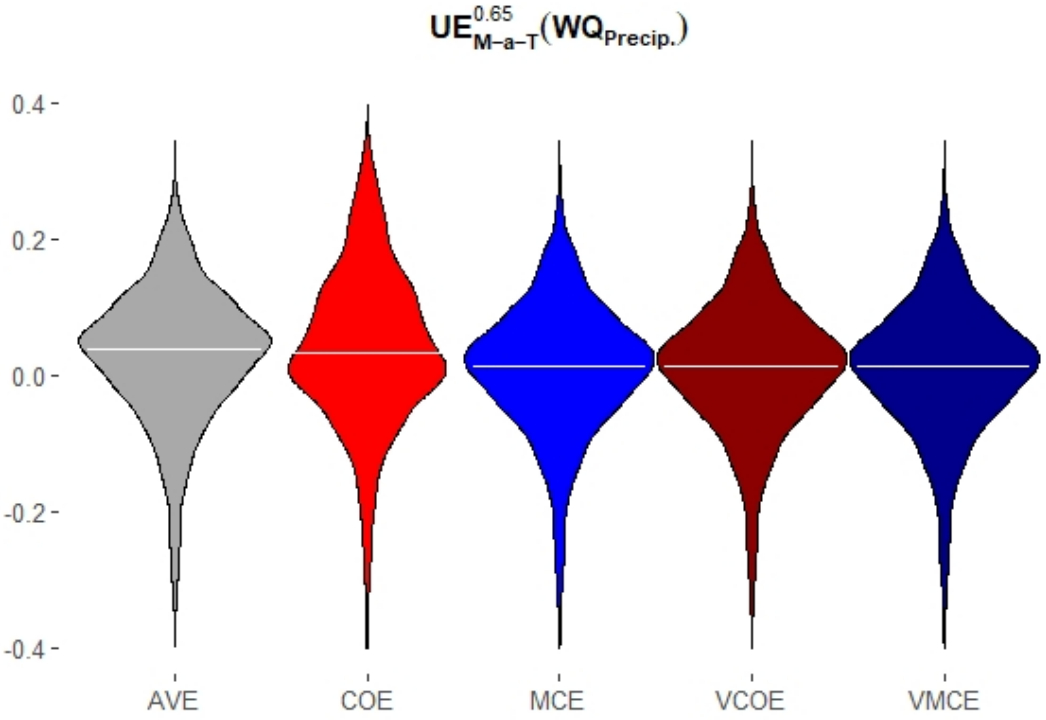

Figure 4.31: Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors with $1-\alpha = 0.65$ in model-as-truth experiments **(years 2021-2100)** for precipitation. The Y-axis is cut at -0.4 and 0.4.

Weighted quantile interval has a much larger spread of uncertainty errors on training and validation periods than prediction interval with $1 - \alpha = 0.65$. The uncertainty error of COE and VCOE method is positively biased across majority of geographical regions. All ensemble weighting methods perform at similar level on training and validation period as well as in model-as-truth experiment when averaged. As $UE^{0.65}(WQ_{Precip.})$ performance is high in model-as-truth experiment but noticeably worse on validation period it indicates that weighted quantile interval is more suitable for far future uncertainty estimation than the prediction interval at $1 - \alpha = 0.65$. We present $WQ^{0.8}$ results for $1 - \alpha = 0.8$ in Figures 4.32, 4.33 and 4.34 below.

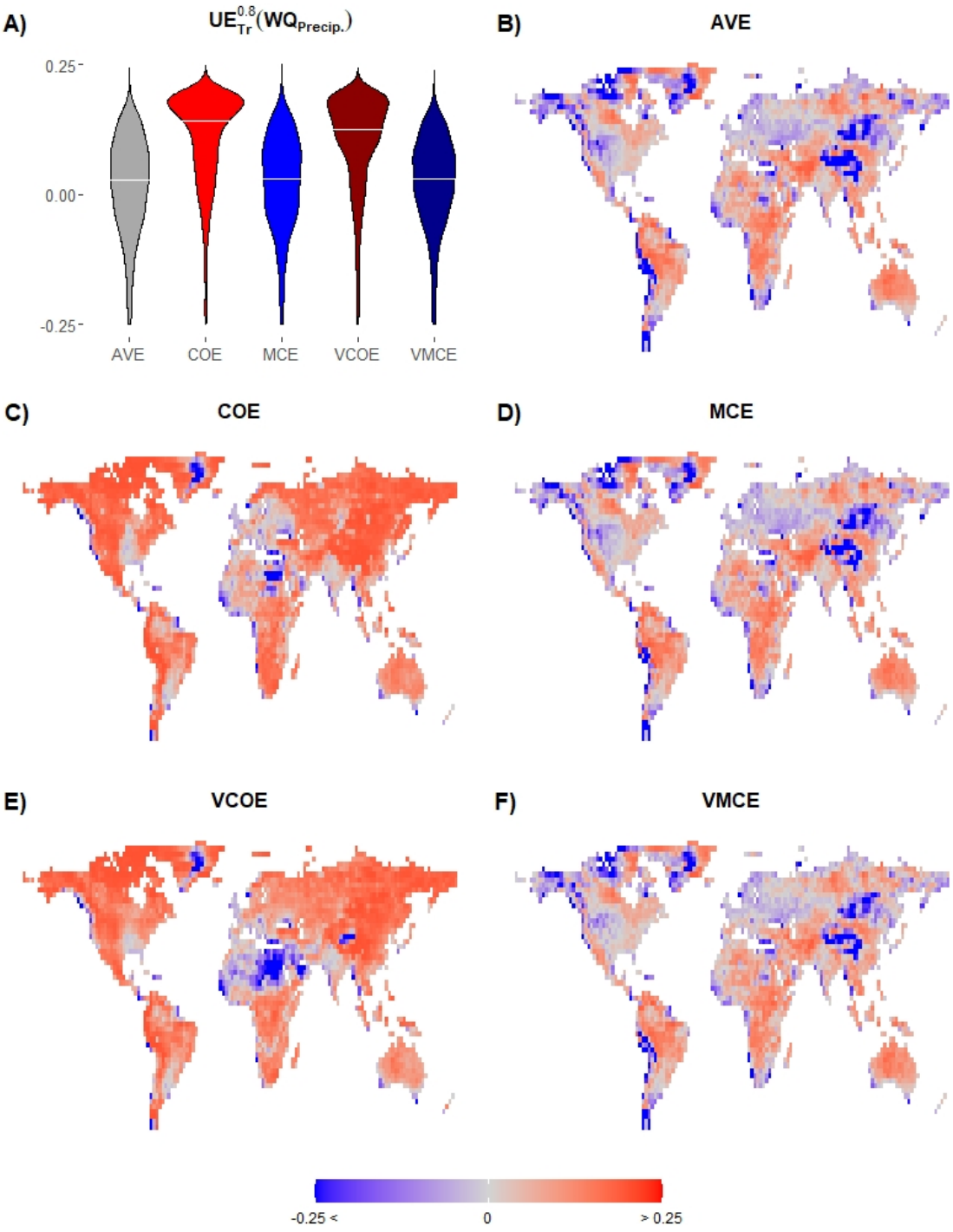

Figure 4.32: Weighted quantile interval uncertainty errors with $1 - \alpha = 0.8$ on training period **(years 1901-1980)** for precipitation. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.25 and 0.25. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors respectively.

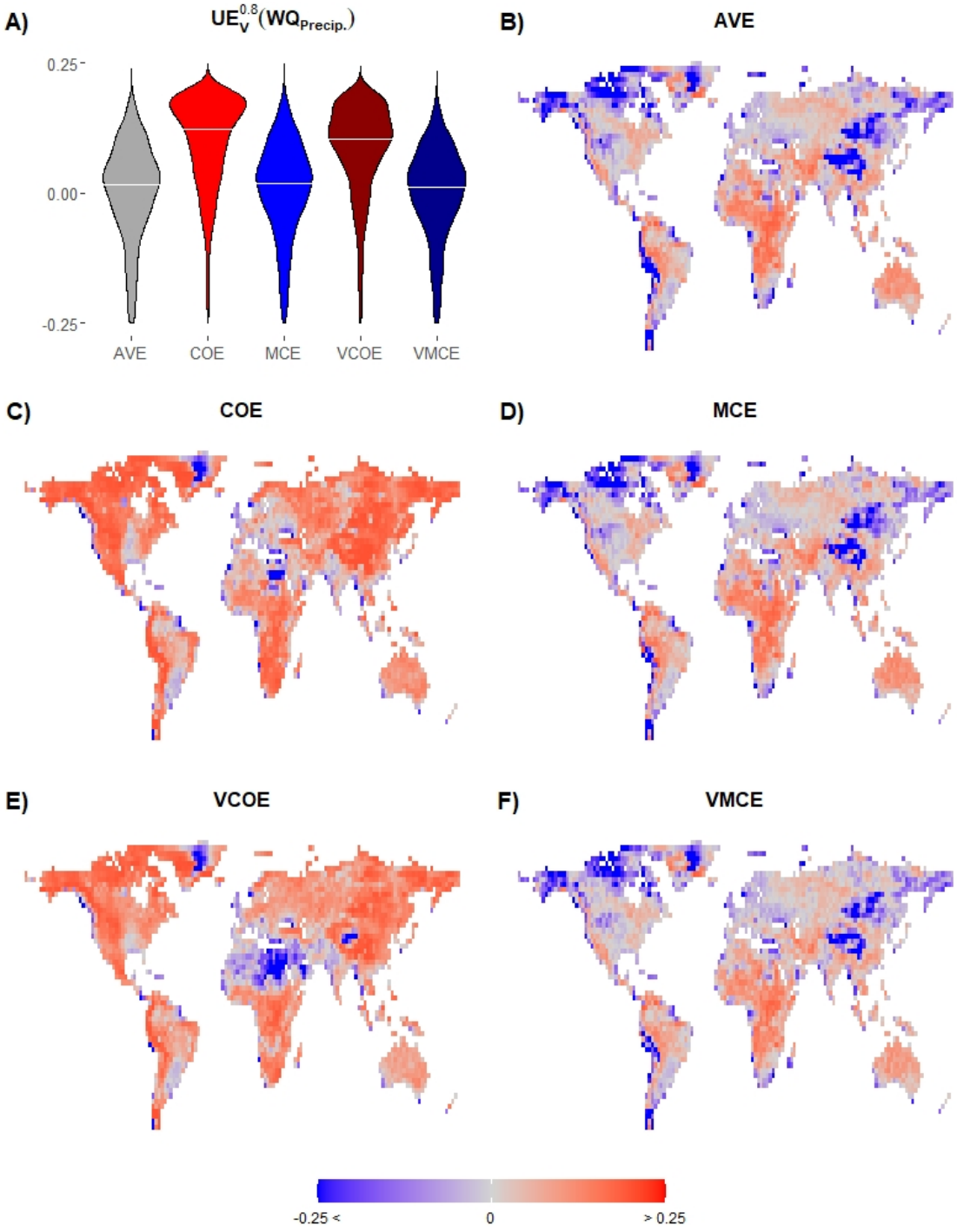

Figure 4.33: Weighted quantile interval uncertainty errors with $1 - \alpha = 0.8$ on validation period **(years 1981-2020)** for precipitation. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.25 and 0.25. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors respectively.

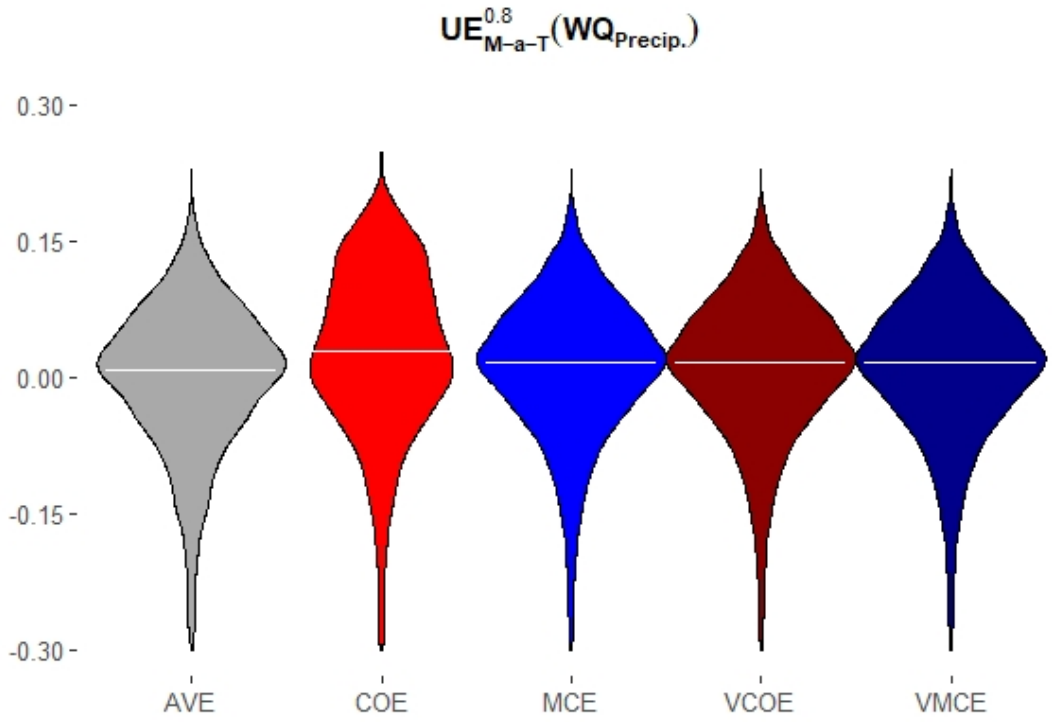

Figure 4.34: Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors with $1-\alpha = 0.8$ in model-as-truth experiments **(years 2021-2100)** for precipitation. The Y-axis is cut at -0.30.

The weighted quantile interval results with $1-\alpha = 0.8$ confirm the findings for weighted quantile interval results with $1-\alpha = 0.65$. The main difference is the scale of the uncertainty error, which is significantly smaller for $UE^{0.8}_{Tr}$, $UE^{0.8}_{V}$ and $UE^{0.8}_{M-a-T}$ compared to $UE^{0.65}_{T}r$, $UE^{0.65}_{V}$ and $UE^{0.65}_{M-a-T}$. As $UE^{0.8}(WQ_{Precip.})$ performance is high in model-as-truth experiment but noticeably worse on validation period it indicates that weighted quantile interval is more suitable for far future uncertainty estimation than the prediction interval at $1-\alpha = 0.8$. We present $WQ^{0.95}$ results for a commonly used $1-\alpha = 0.95$ in Figures 4.35, 4.36 and 4.37 below.

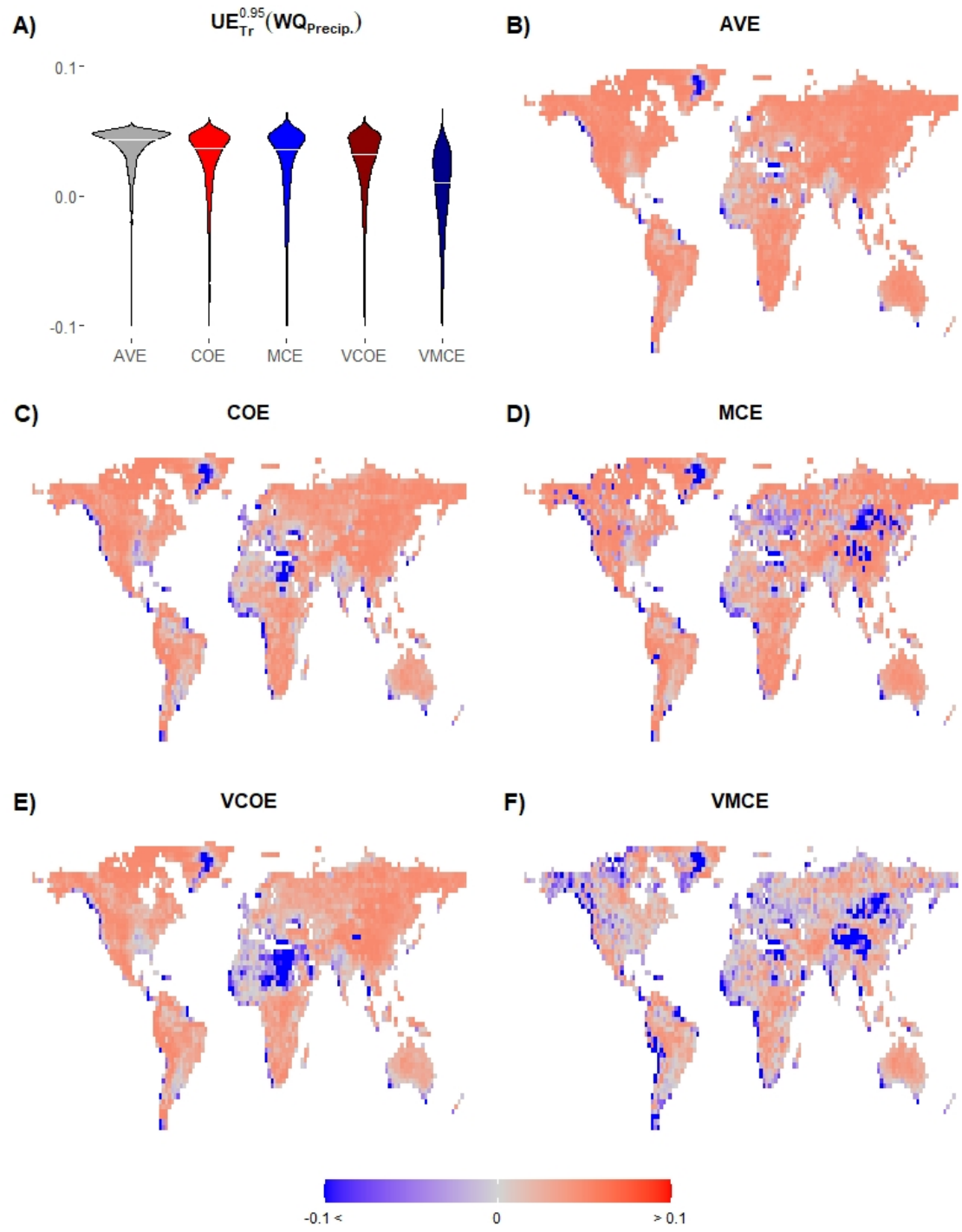

Figure 4.35: Weighted quantile interval uncertainty errors with $1 - \alpha = 0.95$ on training period **(years 1901-1980)** for precipitation. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors using all land points with white lines showing median values. The Y-axis is cut at -0.10. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors respectively.

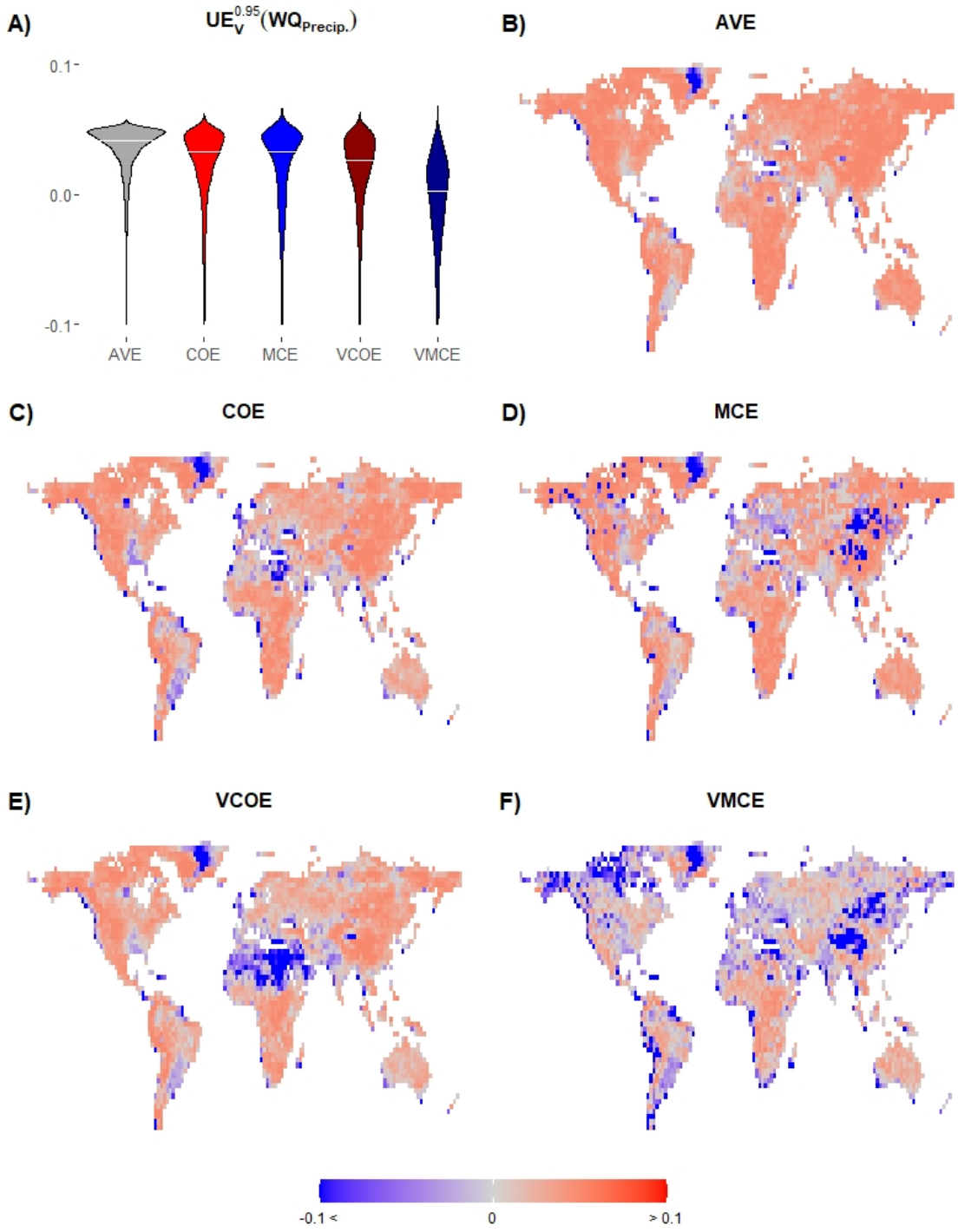

Figure 4.36: Weighted quantile interval uncertainty errors with $1-\alpha = 0.95$ on validation period **(years 1981-2020)** for precipitation. **A)** Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors using all land points with white lines showing median values.The Y-axis is cut at -0.10. **B) - F)** Maps showing geo-spatial distributions of AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors respectively.

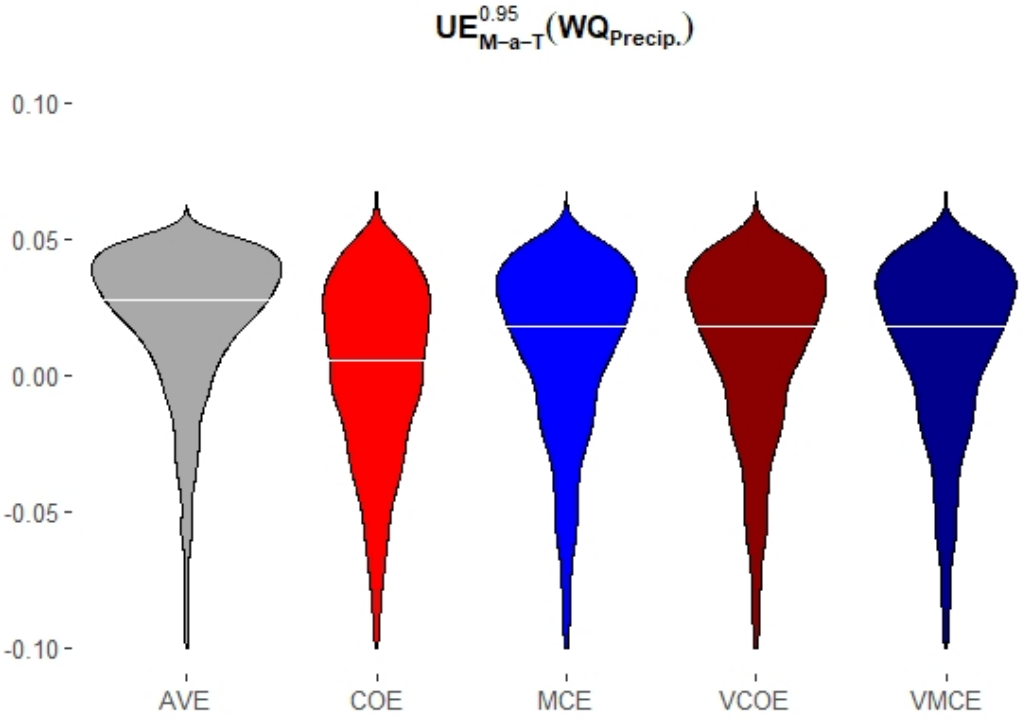

Figure 4.37: Violin plots showing distributions of the AVE, COE, MCE, VCOE, and VMCE methods' weighted quantile interval uncertainty errors with $1-\alpha = 0.95$ in model-as-truth experiments **(years 2021-2100)** for precipitation.The Y-axis is cut at -0.10.

The weighted quantile interval results have $UE^{0.95}(WQ_{Precip.})$ values that are much closer to 0 compared with $UE^{0.65}(WQ_{Precip.})$ and $UE^{0.8}(WQ_{Precip.})$ on both training and validation period as well as in model-as-truth experiment. This confirms the $UE^{0.95}(WQ_{Temp.})$ results obtained in Section 4.3.1. VMCE has the most evenly distributed positive and negative $UE^{0.95}(WQ_{Precip.})$ values for precipitation on training and validation periods. As $UE^{0.95}(WQ_{Precip.})$ performance is high in model-as-truth experiment as well as on validation period it indicates that weighted quantile interval is more suitable for far future uncertainty estimation than the prediction interval at $1-\alpha = 0.95$.

4.3.3 Global climate uncertainty estimation

We summarise all the previous findings in a globally aggregated $PI^{0.95}$ and $WQ^{0.95}$ results for AVE, COE, MCE methods. The aggregation is calculated in four steps. First, we calculate a global monthly time series of observations and weighted ensemble means using all land points as an average of all grid cells' time series weighted by the corresponding grid cell sizes. Then we calculate a global annual time series of observations and weighted ensemble means as an an average of 12 months weighted by the corresponding number of days in each month. We calculate AVE, COE and MCE weighted ensemble means on the obtained global annual data with years 1901-2020 as training period. Finally, we calculate the $PI^{0.95}$ and $WQ^{0.95}$ results on the global weighted ensemble means and compare them to global annual observations.

As the aggregated global annual data is significantly different from geo-spatial monthly data used in Chapter 3 and Chapter 4 so far, this analysis provides one more valuable insight into ensemble weighting methods' performance and difference between prediction and weighted quantile intervals. We present ensemble weigthing methods' uncertainty estimations in pair-wise comparisons between MCE and AVE and between COE and AVE. We present prediction interval results first followed by weighted quantile interval results.

4.3.3.1 Global temperature uncertainty estimation

We present the global annual temperature ensembles' best estimates (AVE, COE, MCE) and their prediction interval ($PI^{0.95}$) with $1 - \alpha = 0.95$ in Figure 4.38 with year 2100 projection results summarised in Table 4.15 below.

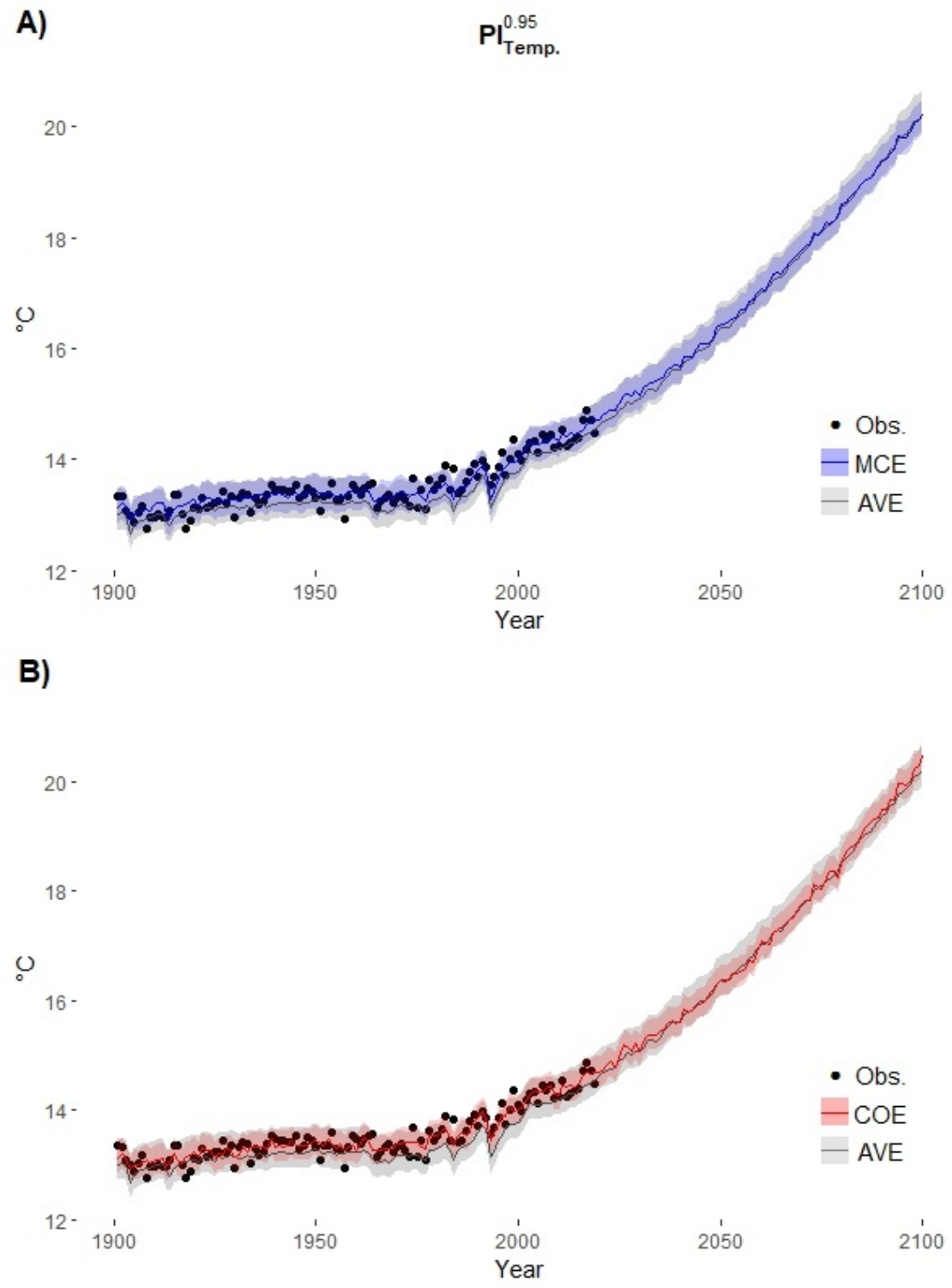

Figure 4.38: Global annual temperature ensembles' best estimates and their prediction interval with $1-\alpha = 0.95$ for years $1901-2100$. Black dots are global annual temperature observations for years $1901-2020$. Grey line & region are AVE results. **A)** Blue line & region - MCE results. **B)** Red line & region - COE results.

Ensemble	BE_{2100}	$LCL_{2100}^{0.95}$	$UCL_{2100}^{0.95}$	$UA_{2100}^{0.95}$
AVE	20.19	19.92	20.66	0.74
COE	20.49	20.23	20.77	0.55
MCE	20.29	19.98	20.61	0.63

Table 4.15: Result comparison of different methods on CMIP6 global annual temperature data using prediction interval. Best Estimate (BE_{2100}) of weighted ensemble means, Lower ($LCL_{2100}^{0.95}$) and Upper ($UCL_{2100}^{0.95}$) Confidence Limits and Uncertainty Area ($UA_{2100}^{0.95}$) values in year 2100 with $1-\alpha = 0.95$. All values are in °C.

All three methods have narrow prediction intervals and the same level of upper control limit in year 2100. The lower control limit in year 2100 is slightly higher for COE than for AVE and MCE. The best estimation are very close for all three methods which confirms the results for temperature data in Chapter 2 (see Table 2.3). Hence, all three methods agree on the projected future temperature in terms of best estimate and prediction intervals. We compare the obtained prediction interval results to the weighted quantile interval results with $1-\alpha = 0.95$ in in Figure 4.39 with year 2100 projection results summarised in Table 4.16 below.

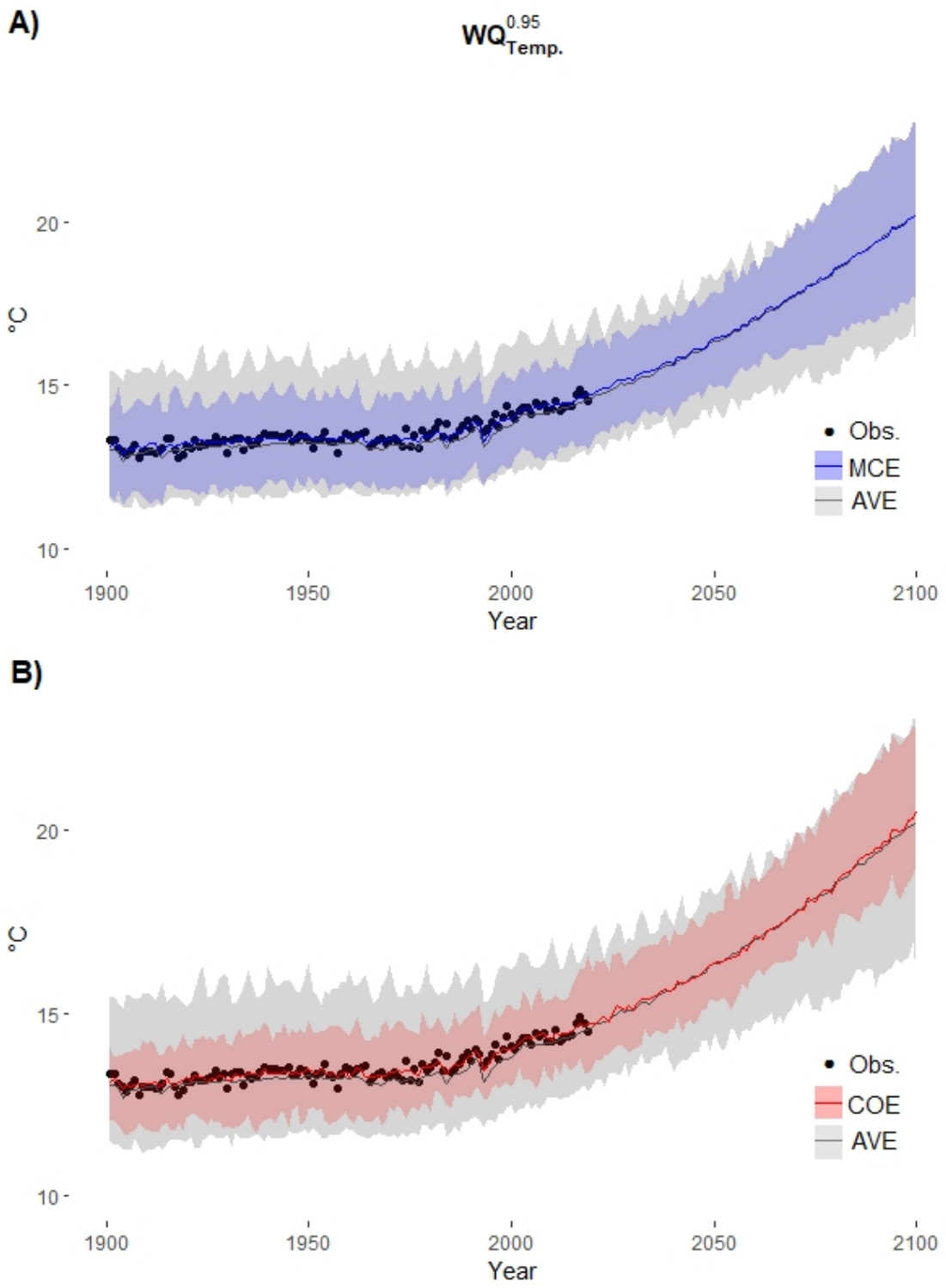

Figure 4.39: Global annual temperature ensembles' best estimates and their weighted quantile interval with $1 - \alpha = 0.95$ for years $1901 - 2100$. Black dots are global annual temperature observations for years $1901 - 2020$. Grey line & region are AVE results. **A)** Blue line & region - MCE results. **B)** Red line & region - COE results.

Ensemble	BE_{2100}	$LCL_{2100}^{0.95}$	$UCL_{2100}^{0.95}$	$UA_{2100}^{0.95}$
AVE	20.19	16.05	23.03	6.98
COE	20.49	19.05	22.83	3.78
MCE	20.29	17.63	22.83	5.19

Table 4.16: Result comparison of different methods on CMIP6 global annual temperature data using weighted quantile interval. Best Estimate (BE_{2100}) of weighted ensemble means, Lower ($LCL_{2100}^{0.95}$) and Upper ($UCL_{2100}^{0.95}$) Confidence Limits and Uncertainty Area ($UA_{2100}^{0.95}$) values in year 2100 with $1-\alpha = 0.95$. All values are in °C.

All three methods have much wider weighted quantile intervals compared to the respective prediction intervals with COE having significantly smaller $UA_{2100}^{0.95}$ than AVE and MCE. The $UCL_{2100}^{0.95}$ using weighted quantile interval is approximately 2 - 2.4 °C higher (depending on ensemble weighting method) than $UCL_{2100}^{0.95}$ using prediction interval. The $LCL_{2100}^{0.95}$ using weighted quantile interval is approximately 1.2 - 3.9 °C lower (depending on ensemble weighting method) than $UCL_{2100}^{0.95}$ using prediction interval.

4.3.3.2 Global precipitation uncertainty estimation

We present the global annual precipitation ensembles' best estimates (AVE, COE, MCE) and their prediction interval ($PI^{0.95}$) with $1-\alpha = 0.95$ in Figure 4.40 with year 2100 projection results summarised in Table 4.17 below.

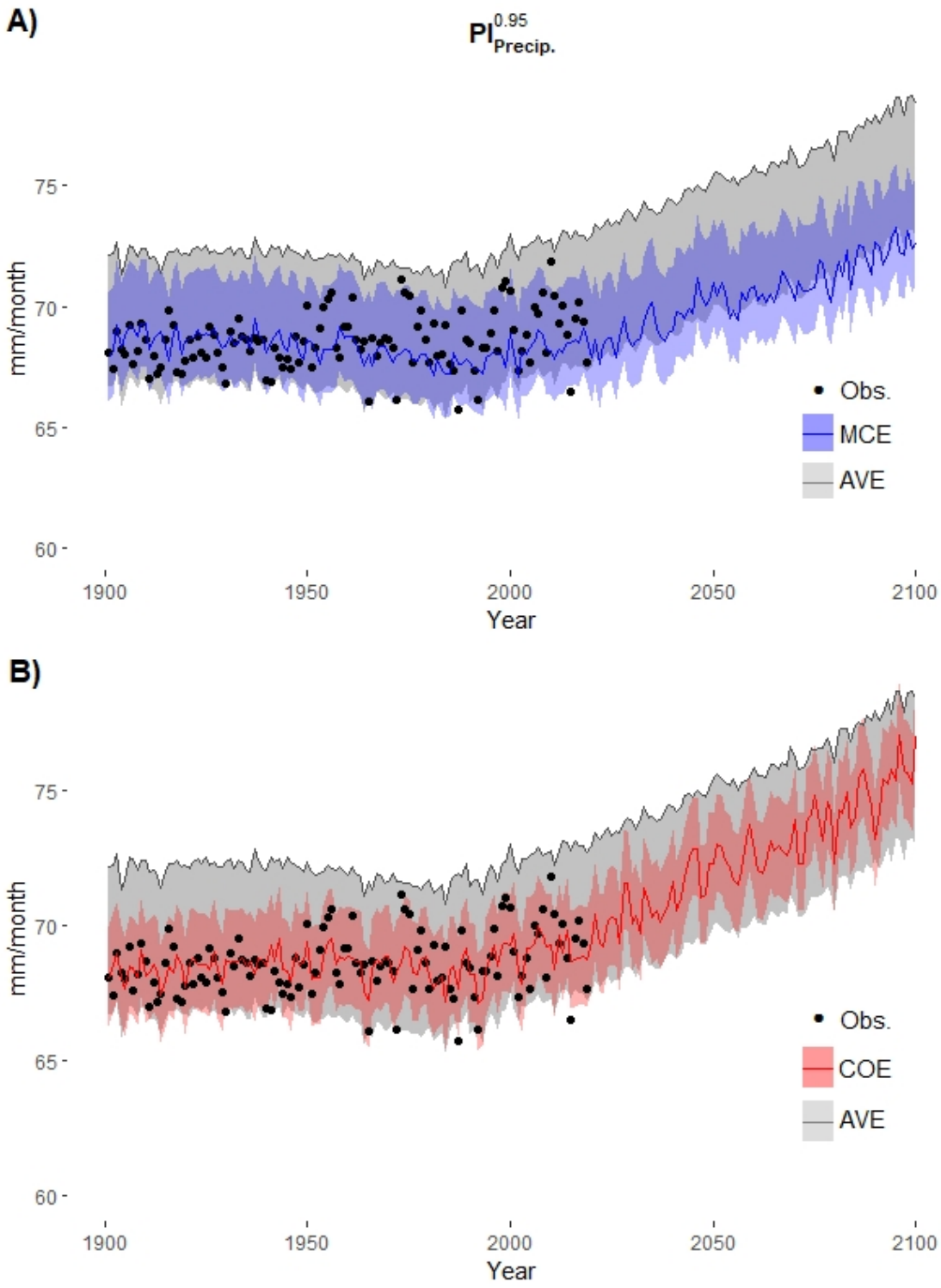

Figure 4.40: Global annual precipitation ensembles' best estimates and their prediction interval with $1-\alpha = 0.95$ for years $1901-2100$. Black dots are global annual temperature observations for years $1901-2020$. Grey line & region are AVE results. **A)** Blue line & region - MCE results. **B)** Red line & region - COE results.

Ensemble	BE_{2100}	$LCL_{2100}^{0.95}$	$UCL_{2100}^{0.95}$	$UA_{2100}^{0.95}$
AVE	78.42	72.95	78.42	5.47
COE	76.98	75.23	78.88	3.64
MCE	72.63	70.84	75.31	4.47

Table 4.17: Result comparison of different methods on CMIP6 global annual precipitation data using prediction interval. Best Estimate (BE_{2100}) of weighted ensemble means, Lower ($LCL_{2100}^{0.95}$) and Upper ($UCL_{2100}^{0.95}$) Confidence Limits and Uncertainty Area ($UA_{2100}^{0.95}$) values in year 2100 with $1-\alpha = 0.95$. All values are in mm/month.

As AVE global annual precipitation mean is higher than the observations due to the aggregation process, the prediction interval becomes a one-sided $1-\alpha$ prediction interval. This demonstrates the flexibility of prediction interval and its applicability to complex data configurations. All three methods have narrow prediction intervals with COE having the lowest value of $UA_{2100}^{0.95}$. The lower control limit in year 2100 is slightly higher for COE than for AVE and MCE. The best estimate and upper control limit in year 2100 is slightly lower for MCE than for AVE and COE. All three methods agree on the projected future precipitation in terms of best estimate and prediction interval estimation (differ less than 10% in BE_{2100}, $LCL_{2100}^{0.95}$ and $UCL_{2100}^{0.95}$ values). We compare the obtained prediction interval results to the weighted quantile interval with $1-\alpha = 0.95$ in in Figure 4.41 with year 2100 projection results summarised in Table 4.18 below.

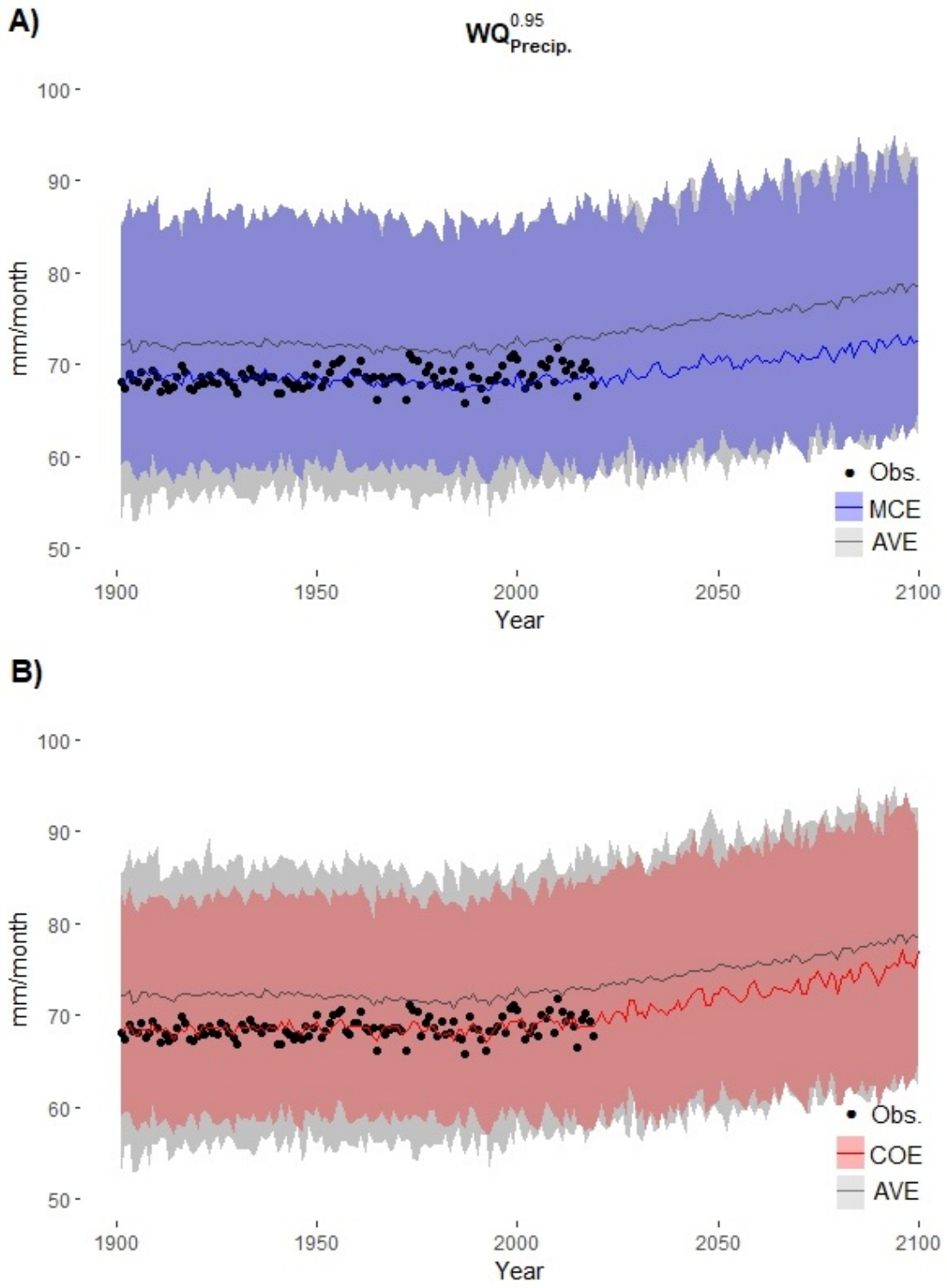

Figure 4.41: Global annual precipitation ensembles' best estimates and their weighted quantile interval with $1 - \alpha = 0.95$ for years $1901 - 2100$. Black dots are global annual temperature observations for years $1901 - 2020$. Grey line & region are AVE results. **A)** Blue line & region - MCE results. **B)** Red line & region - COE results.

Ensemble	BE_{2100}	$LCL_{2100}^{0.95}$	$UCL_{2100}^{0.95}$	$UA_{2100}^{0.95}$
AVE	78.42	62.22	92.78	30.56
COE	76.98	65.13	89.15	24.02
MCE	72.63	65.13	89.74	24.61

Table 4.18: Result comparison of different methods on CMIP6 global annual precipitation data using weighted quantile interval. Best Estimate (BE_{2100}) of weighted ensemble means, Lower ($LCL_{2100}^{0.95}$) and Upper ($UCL_{2100}^{0.95}$) Confidence Limits and Uncertainty Area ($UA_{2100}^{0.95}$) values in year 2100 with $1 - \alpha = 0.95$. All values are in mm/month.

All three methods have much wider weighted quantile intervals compared to prediction intervals with COE having significantly smaller $UA_{2100}^{0.95}$ than AVE and slightly smaller than MCE. The $UCL_{2100}^{0.95}$ using weighted quantile interval is approximately 10.3 - 14.4 mm/-month higher (depending on ensemble weighting method) than $UCL_{2100}^{0.95}$ using prediction interval. The $LCL_{2100}^{0.95}$ using weighted quantile interval is approximately 5.7 - 10.7 mm/-month lower (depending on ensemble weighting method) than $UCL_{2100}^{0.95}$ using prediction interval.

4.4 Discussion

We demonstrated that the prediction interval (PI) uncertainty estimation of future climate is consistent and reliable across different $1 - \alpha$ values, geographical locations and climate data variables. It provides more accurate results than the weighted quantile (WQ) interval for both temperature and precipitation on training and validation data and can be easily applied on data aggregated on geographical and temporary dimensions. However, prediction interval performance is noticeably worse in model-as-truth experiments. This can be attributed to the narrow and constant in time prediction interval despite the ensemble model spread varying with time. This indicates that the prediction interval is less suitable for far future uncertainty estimation than for near future. Weighted quantile interval on the other hand benefits from models being evenly distributed around pseudo-observations (the model chosen as truth) leading to its high performance. In contrast

to the prediction interval, the weighted quantile interval changes if the ensemble model spread changes. This is a limitation of prediction interval approach that needs to be taken into consideration for far future climate projections.

As weighted and unweighted quantile intervals are well-known and accepted uncertainty estimation method in climate sciences (Knutti et al. (2017), Brunner et al. (2020), Lee et al. (2021),etc.) prediction interval can be used as a complimentary approach as it is based on different principles. While weighted and unweighted quantile intervals are dependent on ensemble models' selection and sensitive to ensemble weights, prediction interval does not use that as its input. The prediction interval relies on weighted ensemble means variation and proximity to observations, which are not required for weighted quantiles.

Using the prediction interval (PI) interval results we demonstrated that VMCE method produces optimal results in terms of a combination of UE and UA values on training and validation data. However depending on the goal of uncertainty estimation methods other than VMCE might be preferable (e.g. if only optimal UE performance is taking into consideration, or only negative UE values are of interest). Even in such cases VMCE method is performing at the close to optimal levels and does not have any major disadvantage.

4.5 Conclusion

To verify and extend the results obtained in Chapter 2 and Chapter 3 we introduced and applied a prediction interval (PI) and compared it to a commonly used weighted quantile (WQ) interval in terms of uncertainty error (UE) and uncertainty area (UA) metrics' performance. We compared the aggregated and spatially explicit results for prediction and weighted quantile intervals for different $1-\alpha$ values using AVE, COE, MCE, VCOE and VMCE methods on the CMIP6 data for temperature and precipitation.

The results obtained in Chapter 4 indicate that the introduced prediction interval (PI) is a more consistent, reliable and appropriate in terms of representing $1-\alpha$ level than the

weighted quantile (WQ) interval for near future climate projections but has limitations that need to be taken into consideration for far future projections. Among the five ensemble weighting methods presented in this study VMCE has an optimal uncertainty estimation performance in terms of a combination of uncertainty error (UE) and uncertainty area (UA) metrics on training and validation periods. Based on the above, we can conclude that the prediction interval uncertainty estimation of $VMCE$ method can provide a valuable insight into near future climate state and can be used in conjunction with weighed quantiles interval to provide additional perspectives on future climate states based on different principles of uncertainty estimation.

Chapter 5

Conclusion and future directions

In this thesis we presented a novel approach based on Markov chains to estimate model weights in constructing weighted climate model ensemble means. A comprehensive framework for performance evaluation was presented in the form of multiple, carefully designed metrics (RMSE, climatological monthly RMSE, trend bias, climatology monthly bias and interannual variability), cross-validation procedures (holdout method and model-as-truth experiment), different datasets (KMA, NARCLiM, CMIP5, spatially explicit CMIP6) and climate variables (temperature, precipitation, heatwave amplitude). Three peer methods were selected for comparison – commonly used multi-model mean (AVE), convex optimization (COE) and convex optimisation with varying weights (VCOE). The MCE method and its extension to varying weights – VMCE method were discussed in detail, and their step-wise implementations including mathematical background were presented (Tables 2.2 and 3.2). A new prediction interval approach was introduced to accompany the best estimate of weighted climate model ensemble means with uncertainty estimation. Its step-wise implementation and mathematical background was presented (Table 4.1) and its performance in the form of uncertainty error and uncertainty area was compared to a commonly used weighted quantile interval.

The demonstrated results indicate that applying nonlinear ensemble weighting methods on climate datasets can improve future climate projection in terms of best estimate and

uncertainty estimation accuracy. Even a simple nonlinear structure such as Markov chains shows good performance on different commonly-used datasets compared to linear optimisation approaches. These results are supported by using spatially explicit data, standard performance metrics, cross-validation procedures and model-as-truth performance assessment. The developed MCE and VMCE methods are objective in terms of parameter selection, have a sound theoretical basis and have a relatively low number of limitations. They maintain ensemble diversity, mitigate model interdependence issues and capture some of the nonlinear patterns in the data while optimising ensemble weights. Their comparative performance is shown to be significantly higher on non-Gaussian datasets compared to Gaussian datasets. The prediction interval results demonstrate that it provides consistent, reliable and appropriate estimation of confidence level uncertainty within a calibration/validation framework. Application of the prediction interval to future climate projections where uncertainty varies with time remains questionable though it may provide useful information in addition to methods such as weighted quantiles. Based on the above, we can conclude that MCE and VMCE methods accompanied by prediction interval uncertainty estimation can provide a valuable insight into the future climate state.

Based on the findings presented in this thesis we are confident to suggest future development of nonlinear optimisation methods for weighting climate model ensembles.

List of Figures

List of Tables

Bibliography

Abramowitz, G. and Bishop, C. H.: Climate model dependence and the ensemble dependence transformation of CMIP projections, J. Climate, 28, 2332–2348, https://doi.org/10.1175/JCLI-D-14-00364.1, 2015.

Abramowitz, G., Herger N., Gutmann, E. D., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dynam., 1-20, https://doi.org/10.5194/esd-10-91-2019, 2018.

Almazroui, M., Ashfaq, M., Islam, M.N. et al.: Assessment of CMIP6 performance and projected temperature and precipitation changes over South America, Earth Syst. Environ., 5, 155–183, https://doi.org/10.1007/s41748-021-00233-6, 2021.

Bai, J. and Wang, P.: Conditional Markov chain and its application in economic time series analysis, J. Appl. Econ., 26, 715-734. https://doi.org/10.1002/jae.1140, 2011.

Bishop, C.H. and Abramowitz, G.: Climate model dependence and the replicate Earth paradigm, Clim. Dyn., 41, 885–900, https://doi.org/10.1007/s00382-012-1610-y, 2013.

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., Kushnir, Y., Kimoto, M., Meehl, G. A., Msadek, R., Mueller, W. A., Taylor, K. E., Zwiers, F., Rixen, M., Ruprich-Robert, Y., and Eade, R.: The Decadal Climate Prediction Project (DCPP) contribution to CMIP6, Geosci. Model Dev., 9, 3751–3777, https://doi.org/10.5194/gmd-9-3751-2016, 2016.

Bourdeau-Goulet, S.-C., and Hassanzadeh, E.: Comparisons between CMIP5 and CMIP6 models: Simulations of climate indices influencing food security, infrastructure resilience, and human health in Canada. Earth's Future, 9, https://doi.org/10.1029/2021EF001995, 2021.

Brunner, L., Pendergrass, A., Lehner, F., Merrifield, A., Lorenz, R. and Knutti, R.: Reduced global warming from CMIP6 projections when weighting models by performance and independence, Earth Syst. Dyn., 11, 995-1012, 10.5194/esd-11-995-2020, 2020.

Chao-An, C., Huang-Hsiung, H. and Hsin-Chien, L.: Evaluation and comparison of CMIP6 and CMIP5 model performance in simulating the seasonal extreme precipitation in the Western North Pacific and East Asia, Weather and Climate Extremes, Volume 31, 2021, 100303, ISSN 2212-0947, https://doi.org/10.1016/j.wace.2021.100303, 2021.

Christensen, J. H., Kanikicharla, K. K., Aldrian, E., An, S. I., Albuquerque Cavalcanti, I. F., de Castro, M., Dong, W., Goswami, P., Hall, A., Kanyanga, J. K., Kitoh, A., Kossin, J., Lau, N. C., Renwick, J., Stephenson, D. B., Xie, S. P., Zhou, T., Abraham, L., Ambrizzi, T., ... Zou, L.: Climate phenomena and their relevance for future regional climate change. In Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Vol. 9781107057999, 1217-1308, Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.028, 2013.

Crawford, J., Venkataraman K. and Booth J.: Developing climate model ensembles: A comparative case study, J. Hydrol. , Vol. 568, 160-173, https://doi.org/10.1016/j.jhydrol.2018.10.054, 2019.

Curry, J.: Reasoning about climate uncertainty, Clim. Change, 108, 723, https://doi.org/10.1007/s10584-011-0180-z, 2013.

Del Moral, P. and Penev, S.: Stochastic Processes. From Applications to Theory, Taylor and Francis Group, 2016.

Evans, J. P., Ji, F., Lee, C., Smith, P., Argüeso, D., and Fita, L.: Design of a regional climate modelling projection ensemble experiment – NARCliM, Geosci. Model Dev., 7, 621–629, https://doi.org/10.5194/gmd-7-621-2014, 2014.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Fan, Y., Olson, R., and Evans, J.: A Bayesian posterior predictive framework for weighting ensemble regional climate models, Geosci. Model Dev., 1-22, https://doi.org/10.5194/gmd-2016-291, 2017.

Feng, J., Lee, D., Fu, C., Tang, J., Sato, Y., Kato, H., Mcgregor, J., and Mabuchi, K.: Comparison of four ensemble methods combining regional climate simulations over Asia, Meteorol. Atmos. Phys., 111, 41-53, https://doi.org/10.1007/s00703-010-0115-7, 2011.

Fischer, E. and Schär, C.: Consistent geographical patterns of changes in high-impact European heatwaves, Nat. Geosci., 3, 398–403, https://doi.org/10.1038/ngeo866, 2010.

Fu A, Narasimhan B, Boyd S.L: CVXR: An R package for disciplined convex optimization. J. Stat. Softw., 94(14), 1–34, https://doi.org/10.18637/jss.v094.i14, 2020.

Haughton, N., Abramowitz, G., Pitman, A. and Phipps, S.: Weighting climate model ensembles for mean and variance estimates, Clim. Dynam., 45. 1-13, https://doi.org/10.1007/s00382-015-2531-3, 2015.

Hawkins, E. and Sutton, R: The potential to marrow uncertainty in regional climate predictions, Bull. Am. Meteorol, 90, 1095–1108, https://doi.org/10.1175/2009BAMS2607.1, 2009.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res., 113, https://doi.org/10.1029/2007JD008972, 2008.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B.M.: Selecting a climate model subset to optimize key ensemble properties, Earth Syst. Dynam., 9, 135 - 151. https://doi.org/10.5194/esd-9-135-2018, 2018.

Huang, J.-C., Huang, W.-T., Chu, P.-T., Lee, W.-Y., Pai, H.-P., Chuang, C.-C., and Wu, Y.-W.: Applying a Markov chain for the stock pricing of a novel forecasting model, Commun. Stat. Theory, 46:9, 4388-4402, https://doi.org/10.1080/03610926.2015.1083108, 2017.

Johnson, R. A. and Wichern, D. W.: Applied multivariate statistical analysis. Upper Saddle River, NJ: Prentice Hall, 2002.

Jones, D., Wang, W., and Fawcett, R.: High-quality spatial climate data-sets for Australia, Aust. Meteorol. Ocean., 58, https://doi.org/10.22499/2.5804.003, 2009.

Kharin, V. V., and F. W. Zwiers: Climate predictions with multimodel ensembles, J. Climate, 15, 793–799, https://doi.org/10.1175/1520-0442(2002)015<0793:CPWME>2.0.CO;2, 2002.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in combining projections from multiple climate models, J. Climate, 23, 2739–2758, https://doi.org/10.1175/2009JCLI3361.1, 2010.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, Geophys. Res. Lett., 44, 1909–1918, https://doi.org/10.1002/2016GL072012, 2017.

Krishnamurti, T., Kishtawal, C., LaRow, T., Bachiochi, D., Zhang, Z., Williford, C., Gadgil, S., and Surendran, S.: Improved weather and seasonal climate forecasts from multi-model superensemble, Science, 285, 1548-1550, https://doi.org/10.1126/science.285.5433.1548, 1999.

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., Gadgil, S., and Surendran, S.: Multimodel ensemble forecasts for weather and seasonal climate, J. Climate, 13, 4196–4216, https://doi.org/10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2, 2000.

Lambert, S. and Boer, G.: CMIP1 evaluation and intercomparison of coupled climate models, Clim. Dynam., 17, 83–106, https://doi.org/10.1007/PL00013736, 2001.

Leduc, M., Laprise, R., de Elía, R. and Šeparović, L.: Is institutional democracy a good proxy for model independence?, J. Clim., 29(23), 8301-8316, https://doi.org/10.1175/JCLI-D-15-0761.1, 2016.

Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J.P., Engelbrecht, F., Fischer, E., Fyfe, J.C., Jones, C., Maycock, A., Mutemi, J., Ndiaye, O., Panickal, S., and Zhou, T.: Future global climate: scenario-based projections and near-term Information. In climate change 2021: the physical science basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change(Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 553–672, https://doi.org/10.1175/JCLI-D-15-0761.1, 2021.

Liang, Y., Gillett, N. P., and Monahan, A. H: Climate model projections of 21st century global warming constrained using the observed warming trend. Geophys. Res. Lett., 47, https://doi.org/10.1029/2019GL086757, 2020.

Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., & Knutti, R.: Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. J. Geophys. Res. Atmos., 123, 4509–4526. https://doi.org/10.1029/2017JD027992, 2018.

Majumder, S., Balakrishnan Nair, T. M., Sandhya, K. G., Remya, P. G., and Sirisha, P.: Modification of a linear regression-based multi-model super-ensemble technique and its application in forecasting of wave height during extreme weather conditions, J. Oper. Oceanogr., 11:1, 1-10, https://doi.org/10.1080/1755876X.2018.1438341, 2018.

Masson, D., and Knutti, R. (2011): Climate model genealogy, Geophys. Res. Lett., 38, https://doi.org/10.1029/2011GL046864, 2011.

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., et al.: An updated assessment of near-surface temperature change from 1850: the HadCRUT5 data set. J. Geophys. Res. Atmos., 126, https://doi.org/10.1029/2019JD032361, 2021.

Murphy, J., Sexton, D., Barnett, D., Jones, G., Webb, M., Collins, M. and Stainforth, D.: Quantification of modelling uncertainties in a large ensemble of climate change simulations, Nature, 430, 768-772, https://doi.org/10.1038/nature02771, 2004.

Olson, R., Evans, J., Di Luca, A., and Argueso, D.: The NARCliM project: Model agreement and significance of climate projections, Clim. Res., 69, https://doi.org/10.3354/cr01403, 2016.

Olson, R., An, S.-I., Fan, Y., Chang, W., Evans, J. P., and Lee, J.-Y.: A novel method to test non-exclusive hypotheses applied to Arctic ice projections from dependent models, Nat. Commun., 10(1), 3016, https://doi.org/10.1038/s41467-019-10561-x, 2019.

Pesch, T., Schröders, S., Allelein, H. J., and Hake, J. F.: A new Markov-chain-related statistical approach for modelling synthetic wind power time series, New J. Phys., 17(5), https://doi.org/10.1088/1367-2630/17/5/055001, 2015.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Refaeilzadeh, P., Tang, L. and Liu, H.: Cross-Validation, Encyclopedia of Database Systems, 532–538, https://doi.org/10.1007/978-0-387-39940-9_ 565, 2009.

Sanderson, B. M., Wehner, M., and Knutti, R.: Skill and independence weighting for multi-model assessments, Geosci. Model Dev., 10, 2379-2395, https://doi.org/10.5194/gmd-10-2379-2017, 2017.

Shin, J., Olson, R., and An, S.-I.: Projected heat wave characteristics over the Korean peninsula during the twenty-first century, Asia-Pac. J. of Atmos. Sci., 54, 1-9, https://doi.org/10.1007/s13143-017-0059-7, 2017.

Steinschneider, S., McCrary, R., Mearns, L. O., and Brown, C.: The effects of climate model similarity on probabilistic climate projections and the implications for local, risk-based adaptation planning, Geophys. Res. Lett., 42(12), 5014-5022, https://doi.org/10.1002/2015GL064529, 2015.

Stocker, T., Plattner, G.-K., Dahe, Q.: IPCC climate change 2013: the physical science basis - findings and lessons learned, EGU General Assembly Conference Abstracts, 17003, 2014.

Taylor, K. E., Stouffer, R., and Meehl, G.: An overview of CMIP5 and the experiment design, B. Am. Meteorol. Soc., 93, 485-498, https://doi.org/10.1175/BAMS-D-11-00094.1, 2011.

Xie, S.-P., Deser, C., Vecchi, G. et al.: Towards predictive understanding of regional climate change, Nature Clim. Change 5, 921–930. https://doi.org/10.1038/nclimate2689, 2015.