Quantifying Uncertainty in Multi-model Ensemble Analyses with Conformal Inference

Trevor Harris

August 6, 2024

Texas A&M University → University of Connecticut
Statistics Department

JSM, Portland, OR

Climate models

Climate models are simulations of the Earth's climate system. Include atmosphere, ocean, land, sea ice, and (lately) biogeochemical processes.

"Primary tool for investigating the response of the climate system to changes in forcings (increases in CO2)

... and for making *projections of the future climate*"

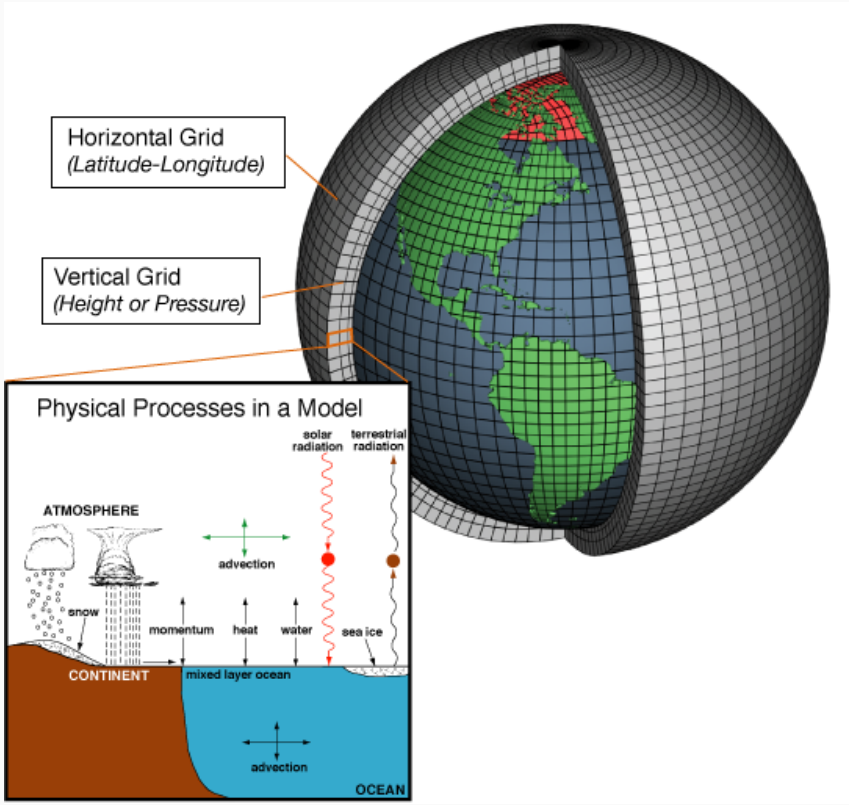

Figure 1: Physical process are resolved over a high resolution grid to generate high resolution model runs. (Climate.gov)

Multi-model Ensemble Analysis

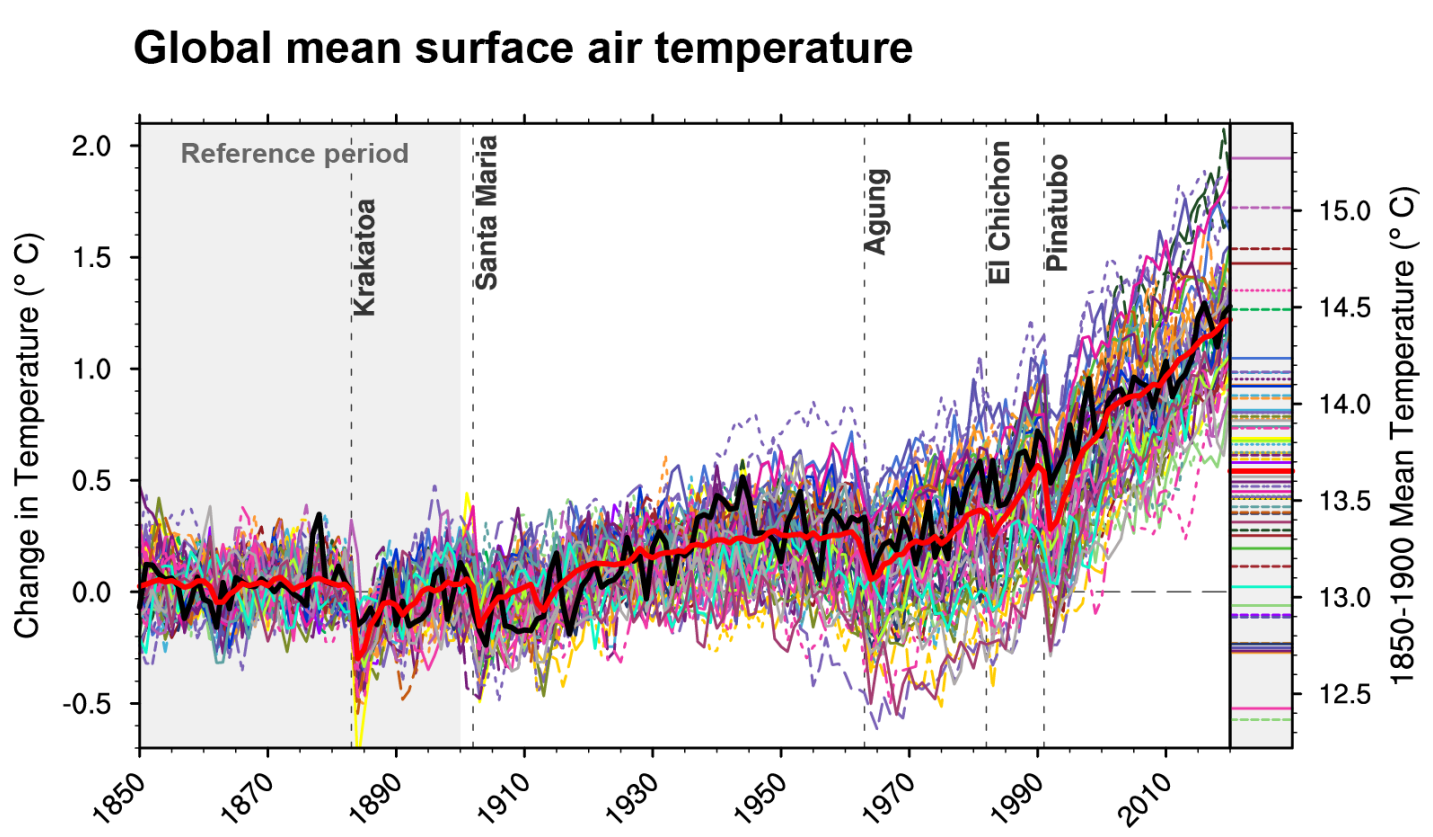

Figure 2: Global mean predictions for each CMIP6 model (colored lines), the model mean (red) and observations (black). Different models yield different predictions.

$\Rightarrow$ Wide variety of models that are being developed. Multi-model ensemble analysis: combine projections and quantify uncertainty.

Multi-model Ensemble Analysis

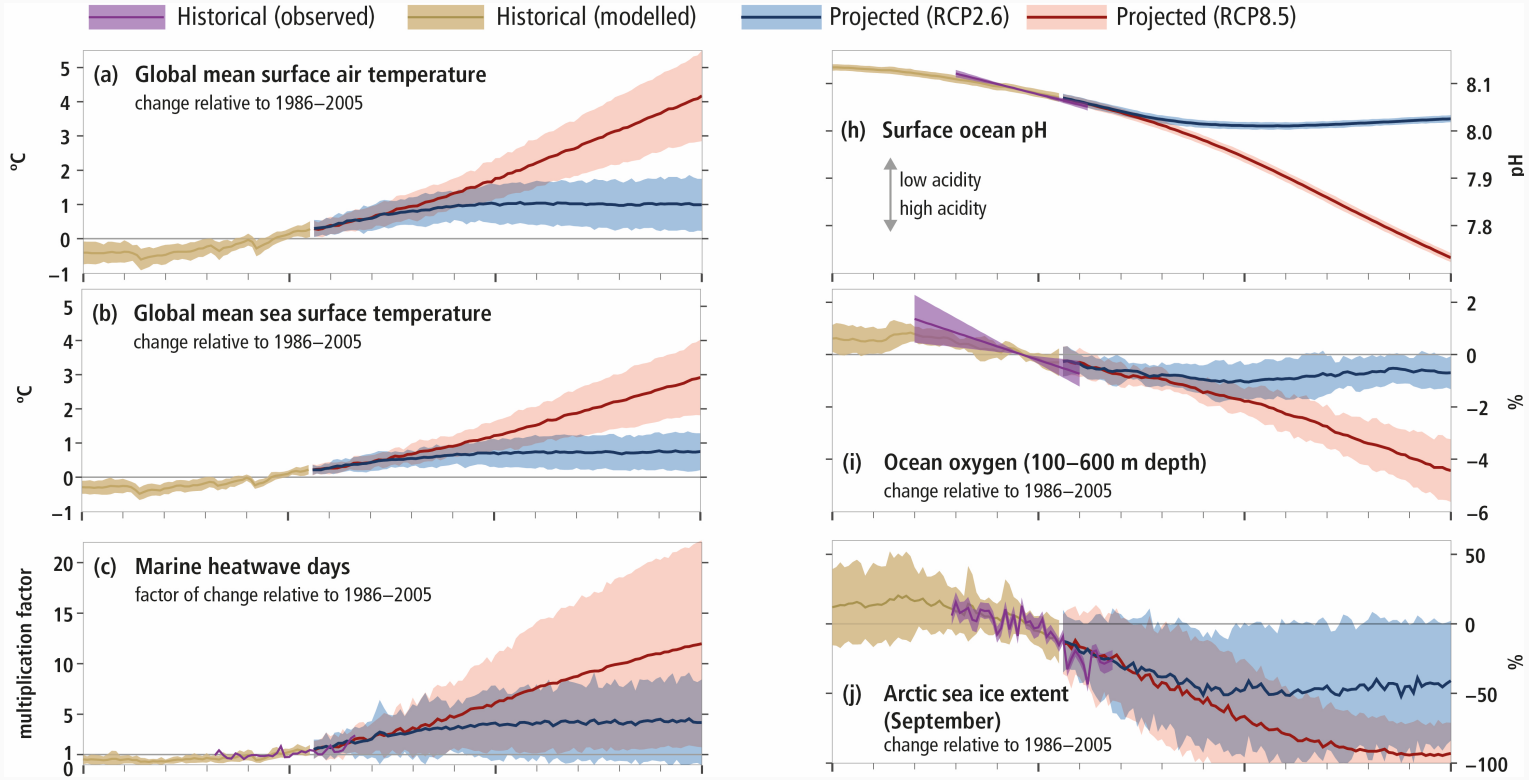

Figure 3: Ensemble uncertainty examples from the Special Report on the Ocean and Cryosphere in a Changing Climate. Bands represent 5–95% model range.

Model summaries (analyses) and inter-model variability (IMV) quantify projections and communicate uncertainty

Previous work – Ensemble Analysis

Multi-model ensemble analysis is a two part problem. (1.) Combine the ensemble (2.) Represent projection uncertainty

There are many approaches to (1.)

- **Ensemble averaging** – democratic and weighted (Giorgi and Mearns, 2002, 2003; Flato et al., 2014; Abramowitz et al., 2019)
- **Bayesian methods** – posterior sampling (Tebaldi et al., 2004; Smith et al., 2009; Bhat et al., 2011; Rougier et al., 2013; Qian et al., 2016; Sansom et al., 2017; Bowman et al., 2018)
- **Regression** – predict observations from models (Räisänen et al., 2010; Bracegirdle and Stephenson, 2012; Ghafarianzadeh and Monteleoni, 2013; Harris et al., 2023)

Regression methods are simple with high skill (GP & CNN). But how to quantify uncertainty?

Previous work – Ensemble Analysis

There are much fewer approaches for (2.)

- **Model uncertainty** – Statistical models (linear models, GPR, etc.) have built-in UQ for prediction intervals.
  - ML models (CNNs, ViTs, etc.) do not.
  - Most methods have have no/poor spatial uncertainty quantification.
- **Inter-model variability** – Add a 90% model spread (re-centered) to the output of the analysis function.
  - Exactly what is done if the analysis is an ensemble average.
  - Does not condition on observational data $\Rightarrow$ high uncertainty
  - Bias correction methods exist to reduce projection uncertainty by conditioning on observations

Can we do a little better?

Given $X \sim \textit{Models}$, $Y \sim \textit{Observations}$, confidence level $\alpha \in (0, 1)$ construct spatially varying prediction regions $C_\alpha(X)$ with confidence guarantees?

Proposal: Conformalize the ensemble

Use **conformal inference** (CI) to construct $C_\alpha(X)$.

- General framework for quantifying uncertainty in the predictions made by arbitrary prediction algorithms. (Vovk et al., 2005; Lei et al., 2018)

  "...can be seen as a method for taking any heuristic notion of uncertainty from any model and converting it to a rigorous one..."

- Does not require asymptotic approximations, priors, correctly specified models X, alignment between models X and observations Y.

- Only requires exchangeability of residuals $Y - \hat{Y}$ over time to construct finite sample valid prediction sets $C_\alpha(X)$.

$$1 - \alpha \leq P(Y \in C_\alpha(X)) < 1 - \alpha + 1/n_{cal}$$

Formalize - model ensemble

- Ensemble of M climate model runs (CMIP6)
  - Historical simulations (1850 – near-present) and future projections (near-present - 2100).
  - Monthly aggregates on 100km to 500km grids.
- $X_{t,i}$ – output of climate model i at time t for $i \in 1, .., M$ and $t \in 1, .., n$. Each $X_{t,i}$ is a gridded field.
- X_t – ensemble of M gridded climate model outputs observed at time t.

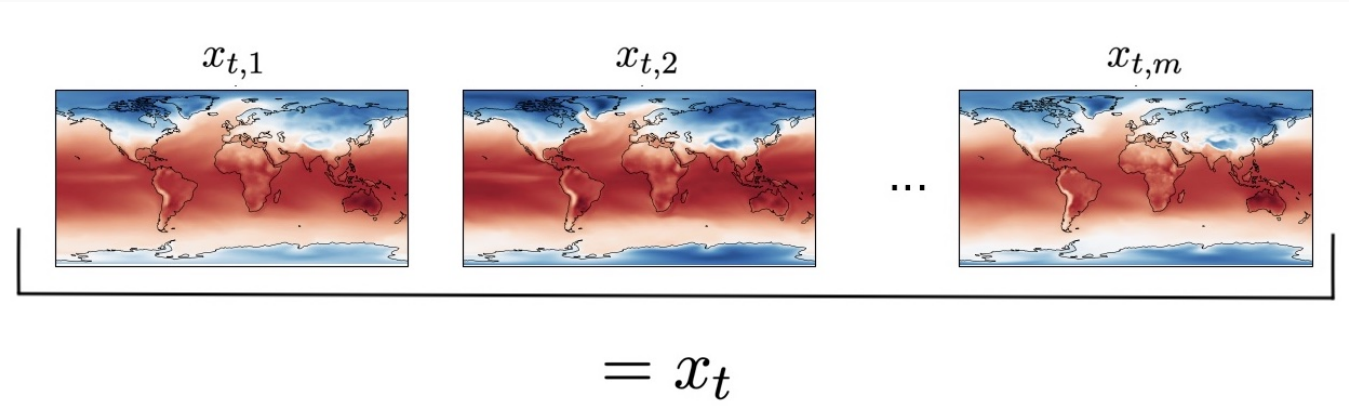

Figure 4: Example model ensemble at time t

Formalize - observations

- Quasi-observational data products – reanalysis fields from ERA5 (ECMWF Reanalysis v5)
  - Spatially complete climate reanalysis (Jan. 1940 - present)
  - Hourly estimates on 31km grid for 137 pressure levels (80km)
  - Monthly aggregates on a single pressure level
- Y_t – reanalysis field at time t. Each Y_t is a gridded field.

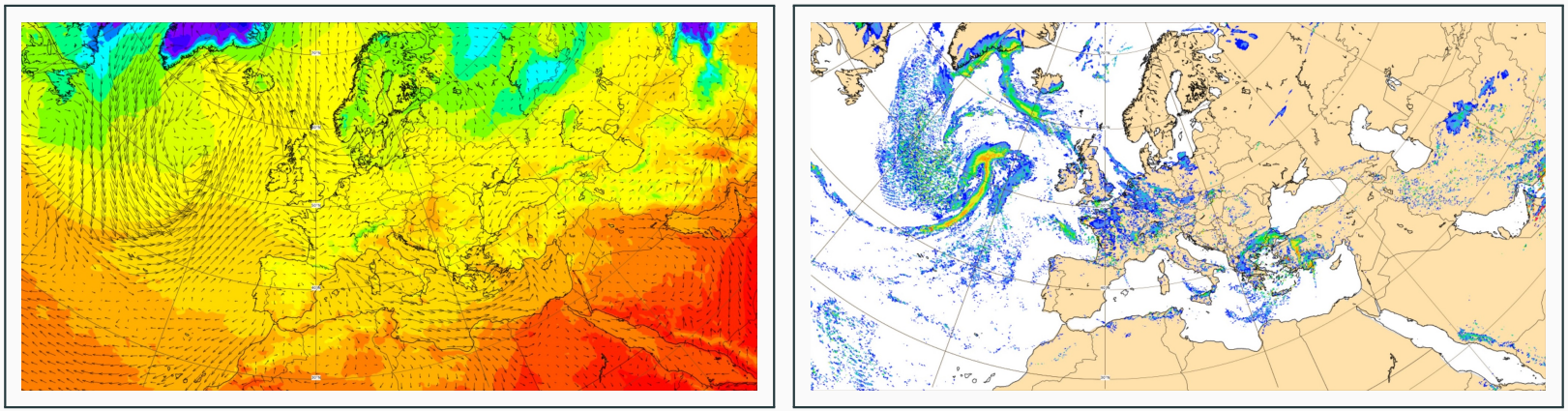

Figure 5: Left: ERA5 Surface temperature and wind speed. **Right:** ERA5 Total Precipitation

Reformulate problem

- Historical dataset $D_{train} = \{(X_t, Y_t)\}_{t=1}^{n}$
  - X_t – ensemble of M gridded climate model outputs observed at time t, run under historical forcings
  - Y_t – reanalysis field at time t
  - Historical: Jan. 1940 - Mar. 2024 (near present)
- Future dataset $D_{test} = \{(X_t, -)\}_{t=n+1}^{N}$
  - X_t – ensemble of M gridded climate model outputs observed at time t, run under future scenario forcings
  - Y_t – not yet observed
  - Future: Apr. 2024 - Dec. 2099

Goal: Given $X_t \sim D_{test}$ and confidence level $\alpha \in (0, 1)$ construct a non-trivial prediction set $C_\alpha(X_t)$ such that

$$P(Y_t \in C_\alpha(X_t)) \geq 1 - \alpha$$

Conformal inference

Conformal inference requires two objects to construct such a $C_\alpha(X_t)$

1. Ensemble analysis function $f_\theta : X \mapsto Y$
2. Scoring function $d : Y \times Y \mapsto \mathbb{R}$

1. $f_\theta(X_t)$ converts the ensemble X_t into a "prediction" of Y_t

$$\text{Estimate:} \quad \hat{\theta} = \arg\min_\theta \mathcal{L}(Y, f_\theta(X))$$

$$\text{Predict:} \quad f_{\hat{\theta}}(X) = \hat{Y}$$

2. $d(Y, \hat{Y})$ ranks out-of-sample residuals, $Y_1 - \hat{Y}_1, ..., Y_N - \hat{Y}_N$ from "most typical" to "most outying" to identify the $(1-\alpha) \times 100\%$ central region, i.e. a $(1-\alpha) \times 100\%$ typical set of residuals.

1. Ensemble analysis functions

- Multi-model analysis model

$$f_\theta : X \mapsto Y$$

  is any statistical/ML/physics based/etc. model of the observational fields conditional on a multi-model ensemble.
  - f_θ can be nearly anything (model average, GLM, GPR, CNN, etc.)
  - Required to takes an ensemble of climate models X_t and return a field $f_\theta(X_t) = \hat{Y}_t \approx Y_t$
  - Prefer f_θ to be as accurate as possible (sharper prediction sets), robust to covariate shift (intrinsic to climate data), and robust to the curse of dimensionality (high dimensional regression).

- Caveat - To get out-of-sample predictions we will have to sample split. Might leave insufficient data for training big ML models.

2. Scoring functions

Data depth defines a natural class of scoring rules for multivariate/functional prediction targets. Depth quantifies how central an observation is with respect to a distribution.

Defn: $D : \mathbb{R}^d \times \mathcal{P} \mapsto [0, 1]$ mapping a data point $z \in \mathbb{R}^d$ and a distribution $P \in \mathcal{P}$ to a real number between 0 and 1 that satisfies

D1 **Translation invariance:** $D(z + b \mid X + b) = D(z \mid X) \quad \forall b \in \mathbb{R}^d, X \sim P$

$\vdots$

D4 **Monotone on rays:** If z^* s.t. $D(z^* \mid X) = \max_{z \in \mathbb{R}^d} D(z \mid X)$, then $\forall r \in \mathbb{S}^d$ the function $\alpha \mapsto D(z^* + \alpha r \mid X)$ decreases ($\alpha > 0$).

D5 **Upper semicontinuous:** The upper level sets $D_\beta(X) = \{z \in \mathbb{R}^d : D(z \mid X) \geq \beta\}$ are closed for all $\beta \in [0, 1]$.

$\Rightarrow$ The more "central" the observation, the higher its depth score, regardless of orientation in space

$\Rightarrow$ Can always find $\beta \in [0, 1]$ such that $P(Y \in D_\beta(X)) \geq 1 - \alpha$ for any coverage level $\alpha \in (0, 1)$

Algorithm – Conformal Central Region

1. Partition the historical data D_{hist} into disjoint training and calibration sets

$$D_{\text{train}} = \{(X_t, Y_t)\}_{t=1}^{n_1} \quad D_{\text{cal}} = \{(X_t, Y_t)\}_{t=n_1+1}^{n},$$

   and let $n_2 = n - n_1$ denote the size of the calibration set.

2. Train the model $f_\theta(\cdot)$ with loss $\mathcal{L}$ on D_{train} as

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \mathcal{L}(f_\theta, D_{\text{train}}).$$

3. Compute $R_t = Y_t - f_{\hat{\theta}}(X_t)$ on D_{cal} and let $\mathcal{R}_{\text{cal}}$ denote the distribution of $R_{n_1+1}, ..., R_n$.

4. Compute the depths of the residual fields R_t with respect to $\mathcal{R}_{\text{cal}}$

$$\mathcal{D}_{cal} = d(R_{n_1+1} \mid \mathcal{R}_{\text{cal}}), ..., d(R_n \mid \mathcal{R}_{\text{cal}}).$$

5. Compute $\tau = Quantile(\mathcal{D}_{cal}, \lceil (n_2 + 1)(1 - \alpha) \rceil)/n_2)$

Conformal Regions v.s. Ensembles

1. **Prediction Region –** τ defines a central region

$$D_\alpha(R) = \{R \sim \mathcal{R}_{\text{cal}} : d(R \mid \mathcal{R}_{\text{cal}}) \geq \tau\}$$

on the residuals $Y - \hat{Y}$. By translation invariance

$$C_\alpha(X_j) = \{f_{\hat{\theta}}(X_j) + R : R \in D_\alpha(R)\}$$

is a prediction central region for any $Y_j \in D_{test}$.

2. **Prediction Ensemble –** Denote the $\lceil (n_2+1)(1-\alpha) \rceil$ least outlying residuals based on the depth function d as

$$R_{(1)}, ..., R_{\lceil (n_2+1)(1-\alpha) \rceil)}.$$

Predict each $Y_j \in D_{test}$ with the set

$$\{f_{\hat{\theta}}(X_j) + R_{(i)}\}_{i=1}^{\lceil (n_2+1)(1-\alpha) \rceil} \subset C_\alpha(X)$$

Conformal Central Region

Theorem 1: Conformal Validity

For a valid relaxed depth function $d(x \mid P)$, the induced central regions $C_\alpha(X)$ have guaranteed coverage

$$P(Y \in C_\alpha(X)) \geq 1 - \alpha$$

and are asymptotically non-conservative in the sample size n_2 of the calibration set D_{cal}

$$P(Y \in C_\alpha(X)) < 1 - \alpha + 1/n_2$$

As long as $R_i = Y_i - \hat{Y}_i$ on D_{cal} and $R_j = Y_j - \hat{Y}_j$ on D_{test} are **exchangeable**, a central region $C_\alpha(X)$ estimated on D_{cal} will also have $(1 - \alpha)$ level coverage on $Y \in D_{test}$.

Numerical experiments

- Test methods ability to quantify uncertainty in future climate projections under different "perfect model" experiments
  - Given M climate model runs, treat one model run as the "truth" (target) and the other $M-1$ as a multi-model ensemble (predictors)
  - Compute out-of-sample coverage ($\alpha = 0.1$) and UQ skill measures on target
  - Repeat for all models to get jackknife-like estimates
- Compare CE (three depth metrics) against IMV and IMV(BC) under six ensemble analysis functions
  - Pointwise average (EA), Weighted pointwise average (WA), the delta method (Delta), Linear model (LM), Gaussian process regression (GP), and a deep convolutional neural network (CNN)
  - ℓ_∞ Depth, Tukey Depth, ℓ_∞ Norm
- 2-meter air surface temperature (TAS: $M = 31$), maximum 2-meter air surface temperature (TMAX: $M = 20$), and Total Precipitation (PR: $M = 30$)

Numerical experiments

- Train on Jan. 1950 - Jul. 2007 data and estimate CE on Jul. 2007 - Mar. 2024 data to match reanalysis availability.
  - Monthly model means. 1011 historical observations. Use the last 200 as a calibration dataset and the first 811 as a proper training set.
  - All models anomalized into forced response (removed monthly means)
- Evaluate all metrics on future simulations (Apr. 2024 - Dec. 2099) based on SSP245 forcings
  - SSP245 – Shared Socioeconomic Pathway 2 with Representative Concentration Pathway (RCP) 4.5 (medium plausible scenario)
  - Coverage and Sliced Wasserstein distance between projection and target
- All models regridded to $90x180$ grid cells (pixels). Done for computational reasons, evidence this improves generalization.

Metrics - size control

Metric	Empirical Coverage ($\alpha = 0.1$)				Average Width (↓)			
Variable	WN	TAS	TMAX	PR	WN	TAS	TMAX	PR
EA. (ℓ_∞)	0.902	0.890	0.893	0.906	6.126	3.717	3.532	2.611
EA. (Tukey)	0.900	0.928	0.929	0.929	6.100	3.717	3.534	2.617
EA. (Norm)	0.901	0.870	0.831	0.906	6.120	3.705	3.524	2.612
WA. (ℓ_∞)	0.900	0.921	0.922	0.910	6.220	3.563	3.474	2.884
WA. (Tukey)	0.900	0.967	0.964	0.938	6.195	3.563	3.472	2.887
WA. (Norm)	0.901	0.933	0.925	0.920	6.213	3.562	3.474	2.885
Delta. (ℓ_∞)	0.902	0.890	0.893	0.906	6.126	3.731	3.605	2.871
Delta. (Tukey)	0.900	0.928	0.929	0.929	6.100	3.731	3.605	2.877
Delta. (Norm)	0.900	0.904	0.893	0.906	6.120	3.730	3.604	2.872
LM. (ℓ_∞)	0.902	0.900	0.913	0.912	6.323	5.341	4.831	3.121
LM. (Tukey)	0.903	0.975	0.969	0.945	6.330	5.346	4.837	3.128
LM. (Norm)	0.901	0.874	0.868	0.903	6.332	5.328	4.816	3.119
GP. (ℓ_∞)	0.902	0.932	0.933	0.908	6.126	3.376	3.295	2.805
GP. (Tukey)	0.900	0.977	0.975	0.949	6.100	3.375	3.293	2.809
GP. (Norm)	0.900	0.931	0.926	0.916	6.120	3.374	3.295	2.806
CNN. (ℓ_∞)	0.900	0.931	0.926	0.916	6.120	3.374	3.295	2.806
CNN. (Tukey)	0.905	0.936	0.935	0.928	6.326	3.687	3.644	2.958
CNN. (Norm)	0.894	0.895	0.882	0.909	6.347	3.684	3.645	2.955

Metrics - size control over time

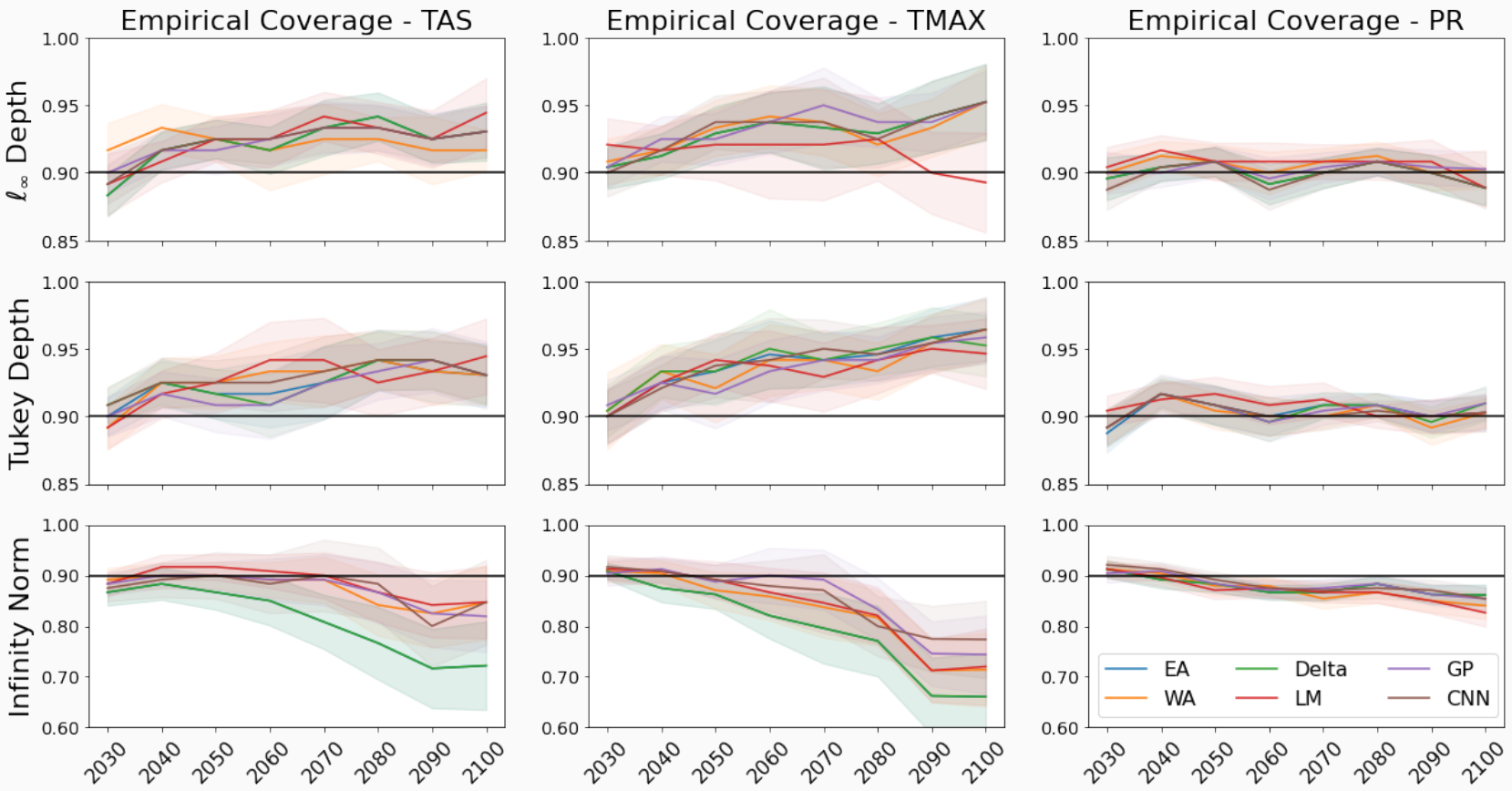

Figure 6: left: Coverage over time or an NN-GPR base using a Conf. Ensemble with raw score quantiles (blue) and adjusted score quantiles (orange). **Right:** CRPS over time for an NN-GPR base using IMV (blue) and a Conf. Ensemble (orange).

Metrics - UQ skill

TAS	EA	WA	Delta	LM	GP	CNN
IMV	1.245 (0.054)	1.277 (0.055)	1.249 (0.055)	1.292 (0.055)	1.288 (0.056)	1.281 (0.057)
IMV (BC)	1.007 (0.067)	1.011 (0.067)	1.008 (0.067)	1.042 (0.067)	1.023 (0.067)	1.014 (0.068)
CE (ℓ_∞)	0.805 (0.043)	0.758 (0.036)	0.793 (0.040)	0.900 (0.033)	0.749 (0.034)	0.739 (0.031)
CE (Tukey)	0.796 (0.041)	0.751 (0.034)	0.784 (0.038)	0.893 (0.031)	0.742 (0.032)	0.730 (0.029)
CE (Norm)	0.801 (0.042)	0.755 (0.037)	0.789 (0.039)	0.899 (0.033)	0.746 (0.034)	0.737 (0.031)
TMAX	EA	WA	Delta	LM	GP	CNN
IMV	1.048 (0.054)	1.066 (0.053)	1.051 (0.054)	1.062 (0.052)	1.068 (0.052)	1.064 (0.053)
IMV (BC)	0.826 (0.054)	0.837 (0.053)	0.827 (0.054)	0.856 (0.051)	0.841 (0.052)	0.837 (0.053)
CE (ℓ_∞)	0.676 (0.041)	0.692 (0.040)	0.675 (0.041)	0.777 (0.035)	0.678 (0.039)	0.678 (0.041)
CE (Tukey)	0.673 (0.039)	0.692 (0.037)	0.672 (0.039)	0.778 (0.033)	0.678 (0.038)	0.676 (0.039)
CE (Norm)	0.675 (0.041)	0.690 (0.039)	0.673 (0.041)	0.775 (0.035)	0.676 (0.040)	0.675 (0.041)
PR	EA	WA	Delta	LM	GP	CNN
IMV	0.222 (0.012)	0.228 (0.012)	0.222 (0.012)	0.230 (0.012)	0.222 (0.012)	0.225 (0.012)
IMV (BC)	0.182 (0.006)	0.186 (0.006)	0.182 (0.006)	0.192 (0.006)	0.184 (0.006)	0.182 (0.006)
CE (ℓ_∞)	0.157 (0.004)	0.161 (0.004)	0.156 (0.004)	0.168 (0.004)	0.157 (0.004)	0.157 (0.004)
CE (Tukey)	0.158 (0.004)	0.162 (0.004)	0.158 (0.004)	0.170 (0.004)	0.158 (0.004)	0.158 (0.004)
CE (Norm)	0.155 (0.003)	0.160 (0.004)	0.155 (0.004)	0.168 (0.004)	0.155 (0.004)	0.156 (0.004)

Table 1: Uncertainty quantification skill metrics for all methods, averaged over projection period (2020-2100), using either the inter-model variability centered at the model predictions (IMV) or the conformal ensemble (Conf.) across the same climatic variables as in table 18.

Metrics - UQ skill over time

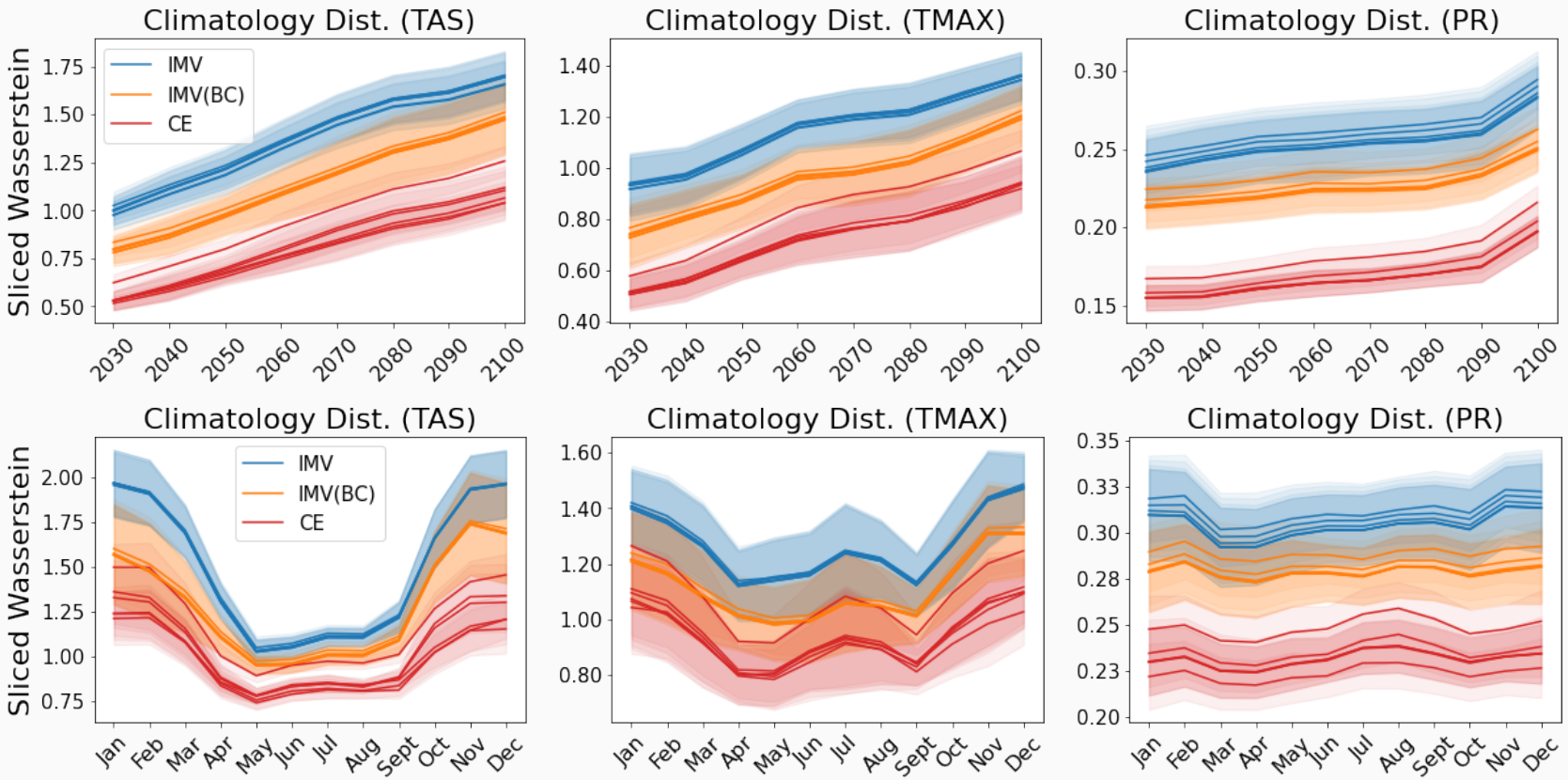

Figure 7: Top: Average sliced Wasserstein (SW) distance between the projections, including all models and UQ methods, and the target within each decade. Solid lines represent the mean SW distance for a given projection over all perfect model experiments and shading represents ± 2 standard errors. **Bottom:** Same as the top row except averaged over months, instead of decades.

Metrics - UQ skill over space

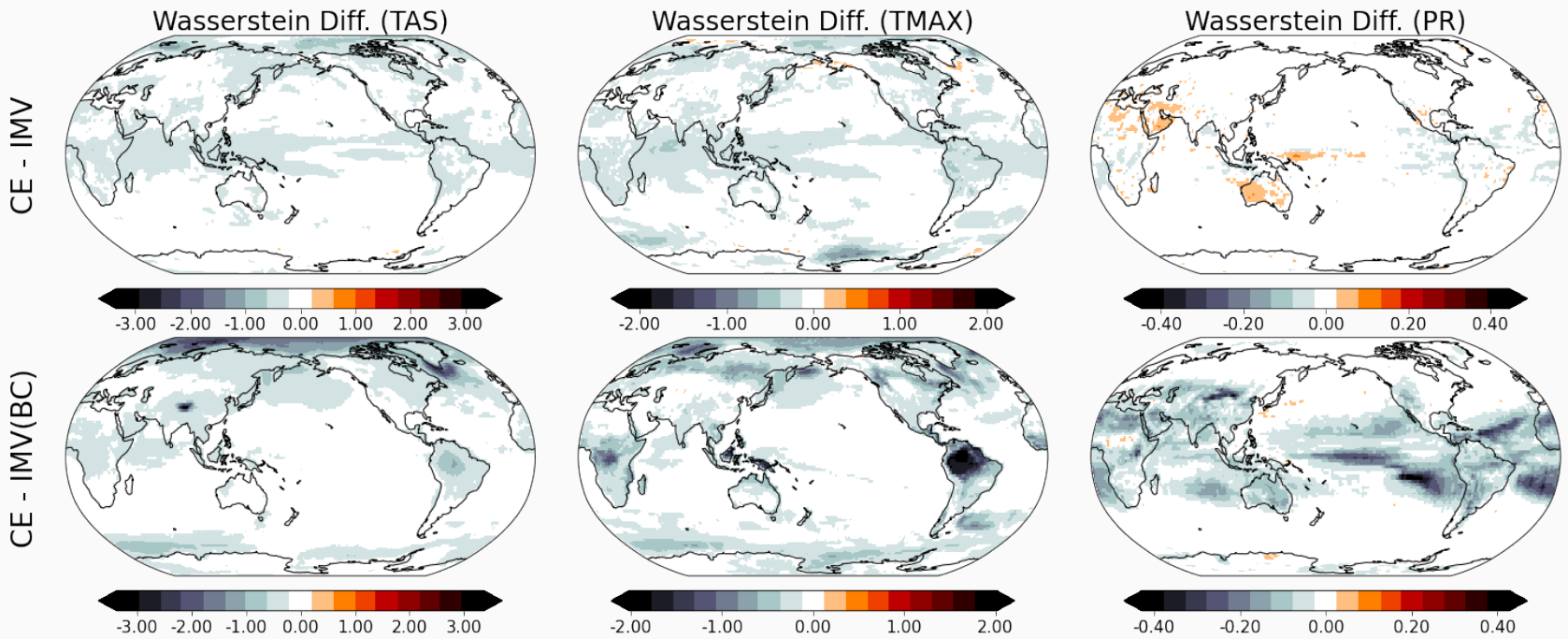

Figure 8: Top: Average difference between the Wasserstein distance of all models using a conformal ensemble (CE) and all models using inter-model variability (IMV) to quantify uncertainty. Wasserstein distances are computed over the test period 2024-2100. **Bottom:** Same as the top row, except computed against all models using bias corrected inter-model variability (IMV(BC)).

Metrics - calibration sensitivity

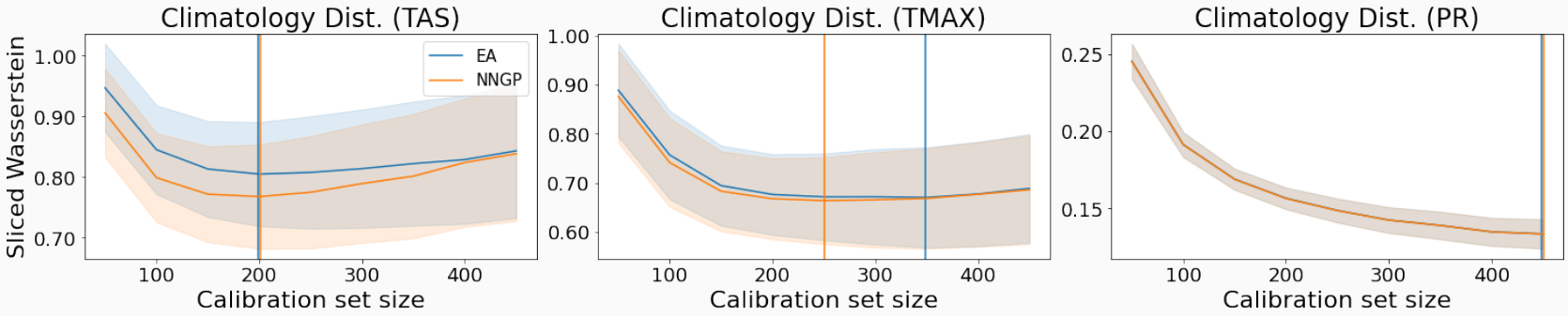

Figure 9: Blue lines show the mean, over all perfect model experiments, sliced Wasserstein distance between EA + CE and the target model over the test period (2024-2100) using an increasing large calibration set size. Orange line show the same except using GP + CE. Blue and orange shading represent ± 2 standard errors from the mean.

Choosing the calibration set size depends on the variable. Variables with less covariate shift (PR) benefit from larger calibration sets. TAS and TMAX exhibit more covariate shift.

Application - Global multi-model analysis

Goal — Combine future simulations (under SSP245) from a multi-model ensemble into a single projection with UQ

- Project 2-meter surface temperature (TAS) and Total Precipitation (PR) monthly averages under SSP2-45
- Use the 31 and 30 member model ensemble from the experiments as input. Predict the corresponding ERA5 reanalysis fields.
- Compare projections against IMV and IMV(BC)
- All models and reanalysis anomalized into forced response (removed monthly means)

1. Validation - Train on (Jan. 1940 - June 2007), estimate CE on (Jul. 2007 - Nov. 2015), evaluate on (Dec. 2015 - Mar. 2024).
2. Quantile projections - compare projected 95% quantile range of CE, IMV, IMV(BC)
3. Global averages - compare global averages under a range of models

Validation

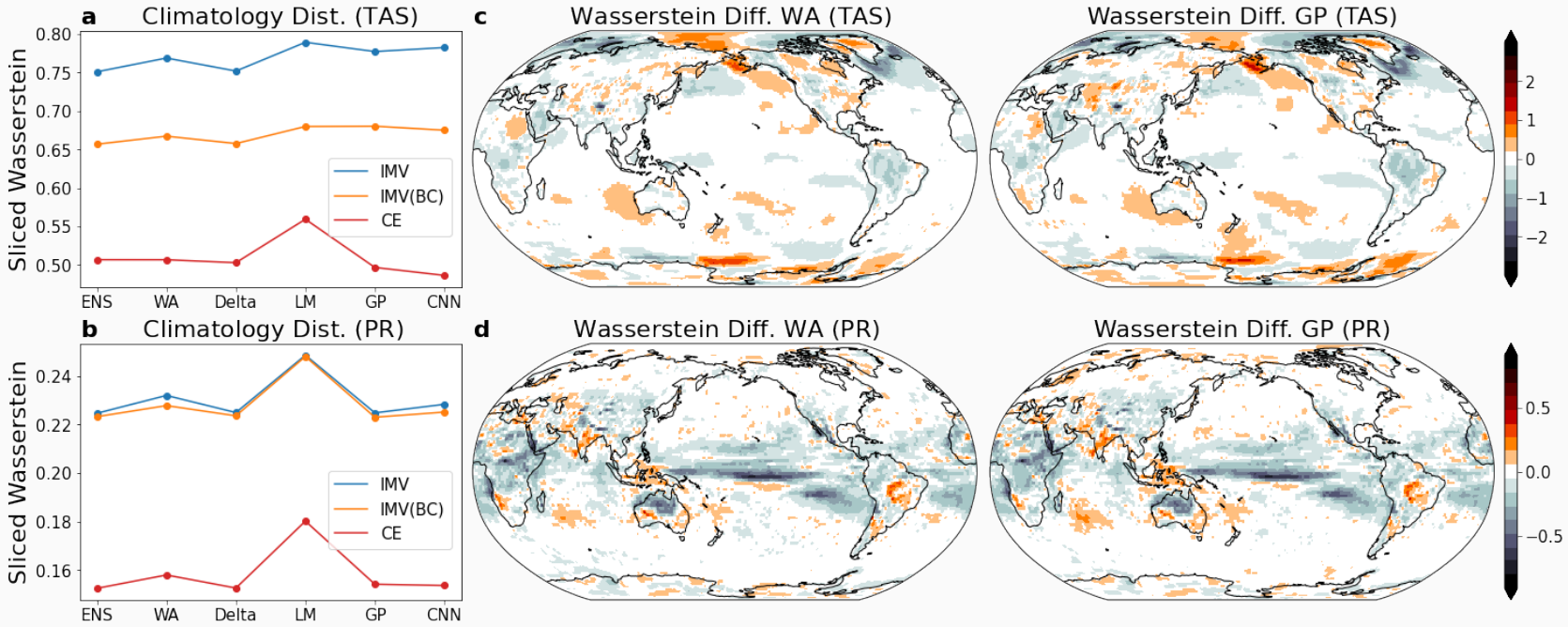

Figure 10: Panels (**a**) and (**b**) show the sliced Wasserstein distance computed between the projected distribution, for each model plus UQ combination, and the reanalysis data on the held out dataset for TAS and PR, respectively. Panels (**c**) and (**d**) show pointwise Wasserstein difference between either a weighted average (WA) model with a CE or a Gaussian process (GP) model with a CE and the ensemble average (EA) using IMV. Red areas indicate that EA + IMV has a lower Wasserstein distance to the observations than WA + CE (or GP + CE), and grey areas indicate the reverse.

TAS - quantile projection

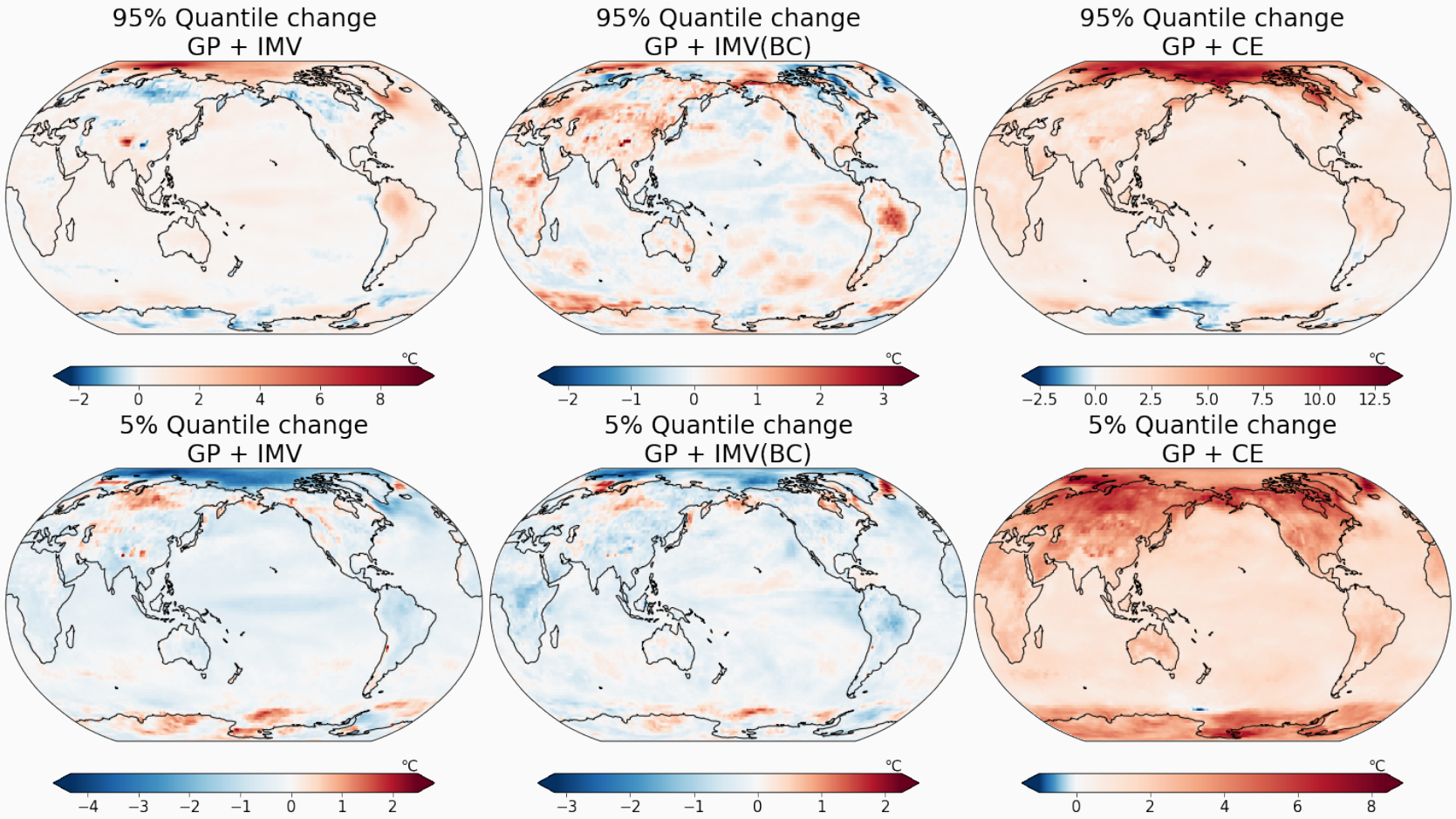

Figure 11: Top: Pointwise 95% quantile difference maps for GP with each UQ method. Maps show the difference between the projected 95% quantile for TAS over the period 2024-2054 and the historical 95% quantile for TAS over the period 1960-1990 at that location. **Bottom:** Same as the top row but using 5% quantiles instead.

PR - quantile projection

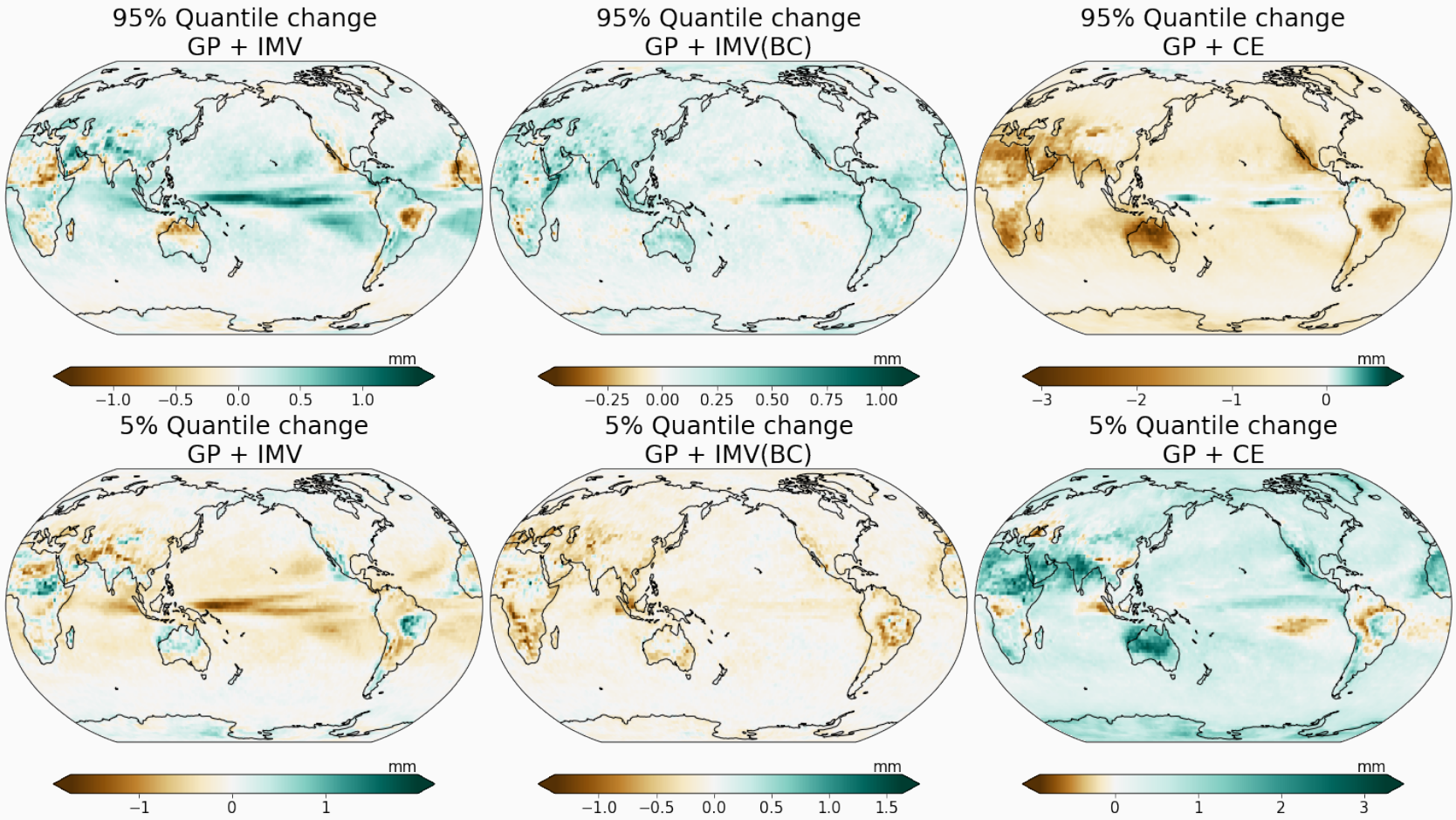

Figure 12: Top: Pointwise 95% quantile difference maps for GP with each UQ method. Maps show the difference between the projected 95% quantile for PR over the period 2024-2054 and the historical 95% quantile for PR over the period 1960-1990 at that location. **Bottom:** Same as the top row but using 5% quantiles instead.

Global projection

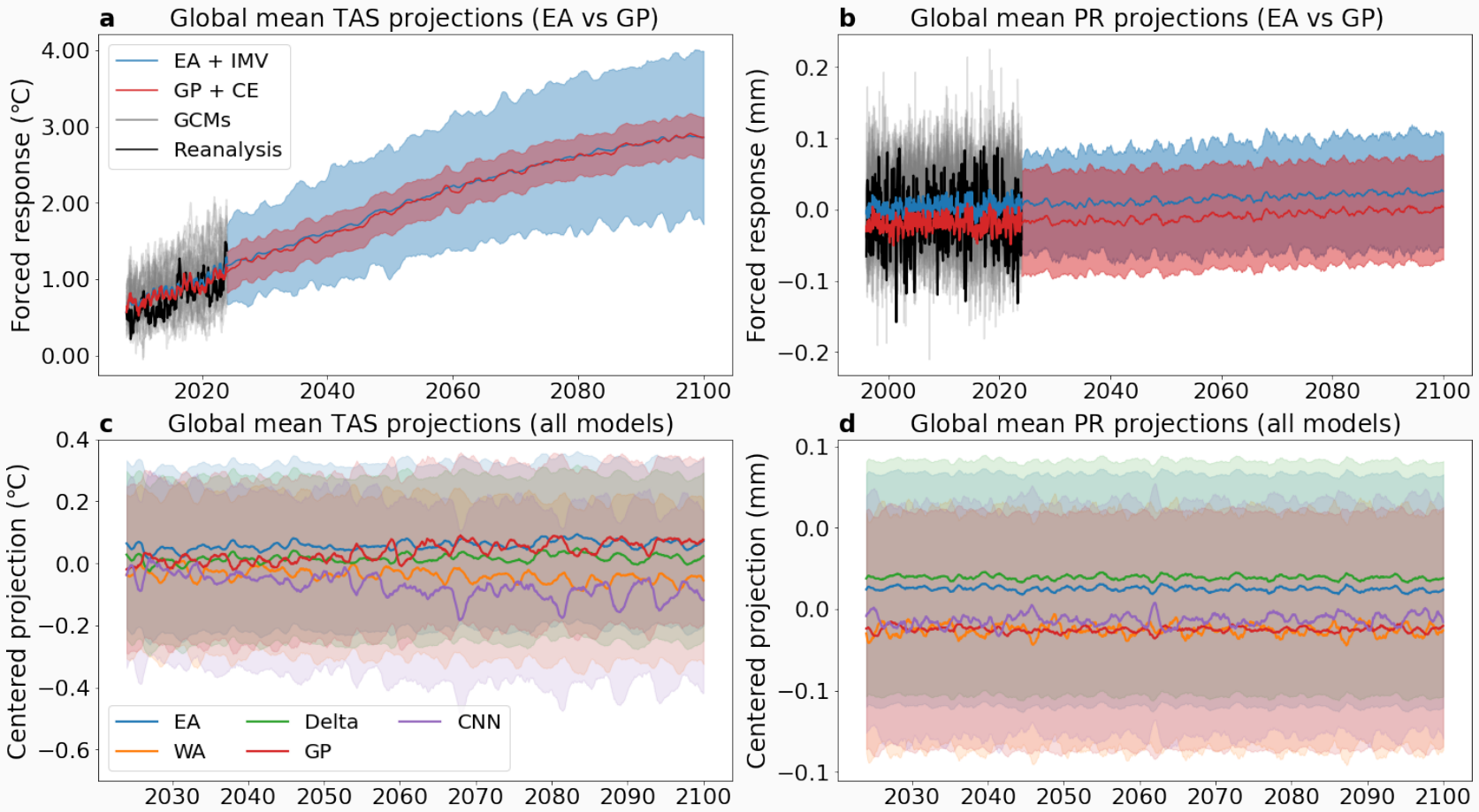

Figure 13: Panels (a) and (b) show the projected global mean monthly averages using EA + IMV and GP + CE. Panels (c) and (d) show the de-meaned projections, with CE based bands, for all models, excluding LM. Projections and uncertainty bands are smoothed using a 12 month moving average to reduce seasonal variability in the plot.

Summary

- Introduced conformal inference as a tool for quantifying uncertainty in climate projections. Conditions UQ on historical observations to reduce projection uncertainty.
  - Compliments traditional model spread and model based UQ
  - Works for any ensemble analysis function and (proper) depth function. Improves with ensemble analysis function skill.
- Validated our approach on a wide variety of model experiments
  - Method has exact coverage (under optimal settings), with mild over/under coverage depending on the adjustment in realistic settings
  - Generally improves UQ skill as measured by Wasserstein distance
  - Robust over time, but falls short of IMV and IMV (BC) marginally at a few critical locations.
- Application to reanalysis data shows
  - More concentrated projections than the IMV in TAS, similar projections as IMV in PR.
  - Little difference between ensemble analysis functions except for small biases.

Limitations - i.e. future work

- Does not account for traditional uncertainty sources:
  - model uncertainty, model inadequacy, natural variability, internal variability, parameter uncertainty, etc.
  - Conformal projection sets do not "expand" with increased data uncertainty and increasing time horizon
  - Hybrid methods?
- Limited ability to model tails due to finite calibration set sizes
  - Currently no way to sample conformal sets to explore the tails
  - Importance sampling?
- Ensemble members may not be physically plausible
  - Prediction sets might miss important regional features.
  - Physically constrained models? PINNs?
- Should not be trusted when exchangeability breaks down – long range projection sets should always be suspect!

Reference

Harris, T. & Sriver, R. (2024). *Quantifying uncertainty in ensemble analyses with conformal inference*. Submitted to the Annals of Applied Statistics.