Quantifying Uncertainty in Multi-model Ensemble Analyses

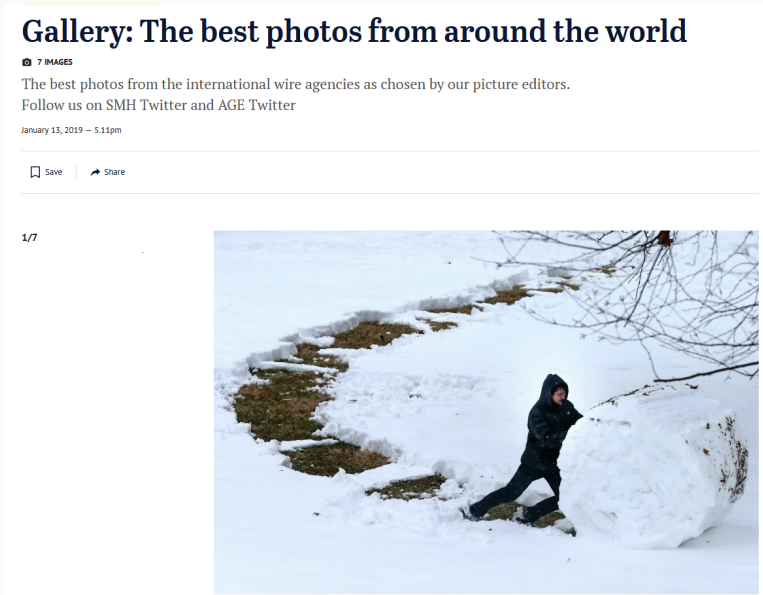
Trevor Harris

March 7, 2024

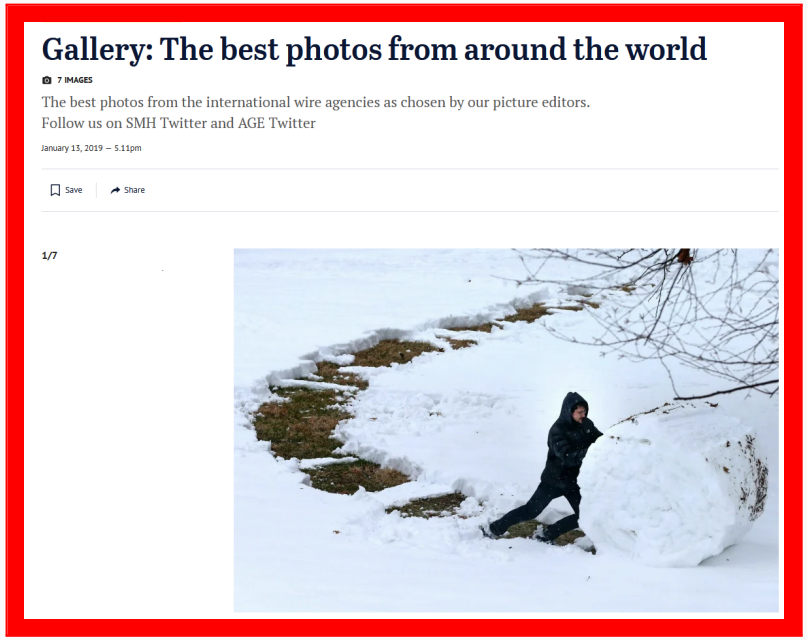
Texas A&M University

Brigham Young University, Provo, UT

Two truths and a lie



Two truths and a lie



Climate models

Climate models are mathematical models of how energy and matter interact in the ocean, atmosphere, and land.

"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*"

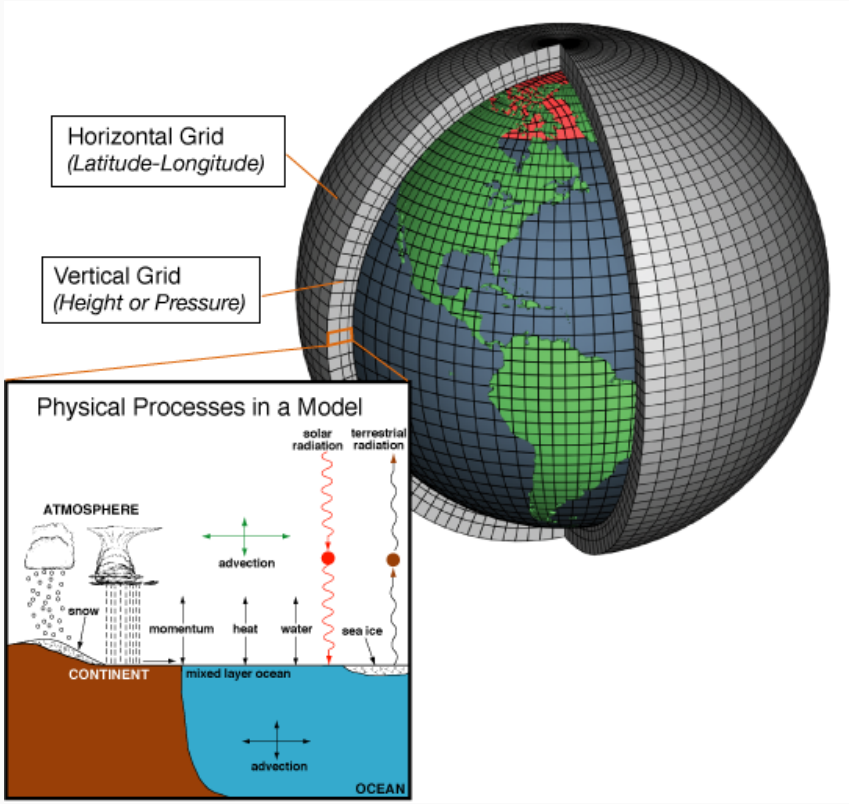


Figure 1: The advanced equations are based on the fundamental laws of physics, fluid motion, and chemistry, solved over a high resolution grid. (Climate.gov)

Multi-model Ensemble Analysis

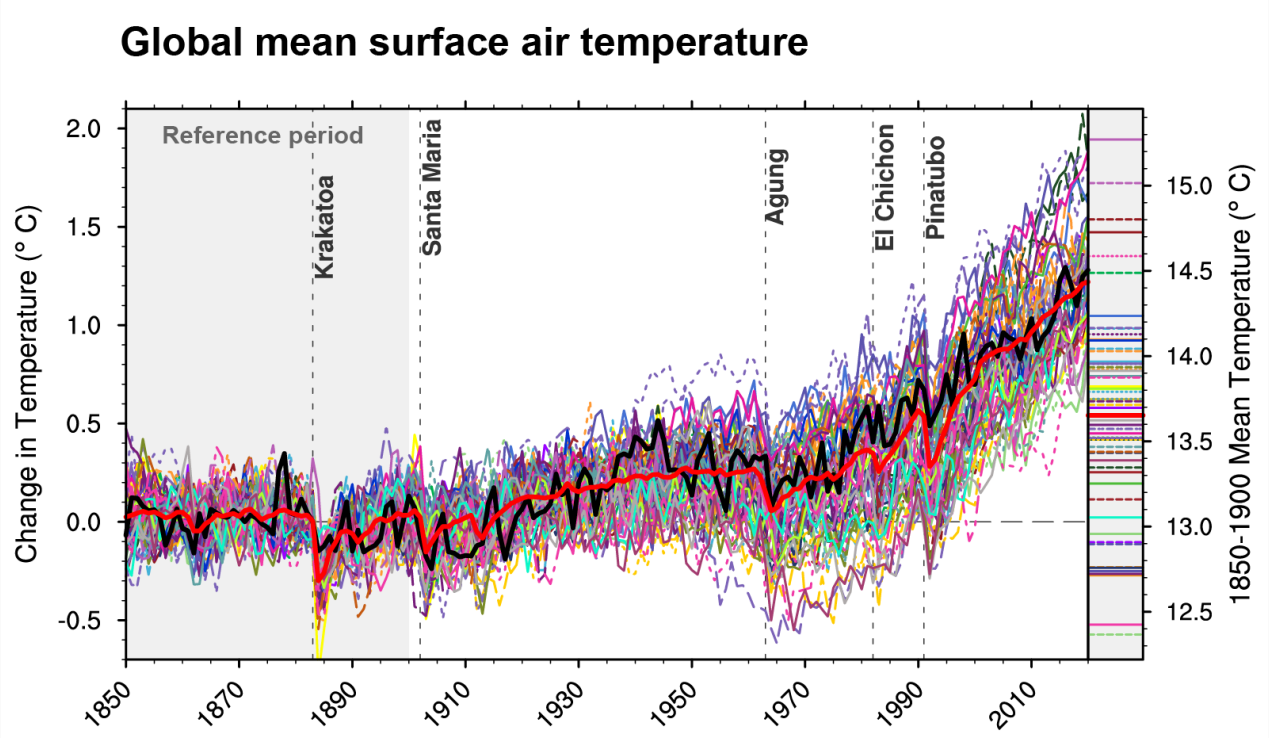


Figure 2: Global mean predictions for each CMIP6 model (colored lines), the model mean (red) and observations (black). Different models yield different predictions.

Multi-model ensembles – output of many different climate models

Ensemble analysis – combined estimate with uncertainty

Multi-model Ensemble Analysis

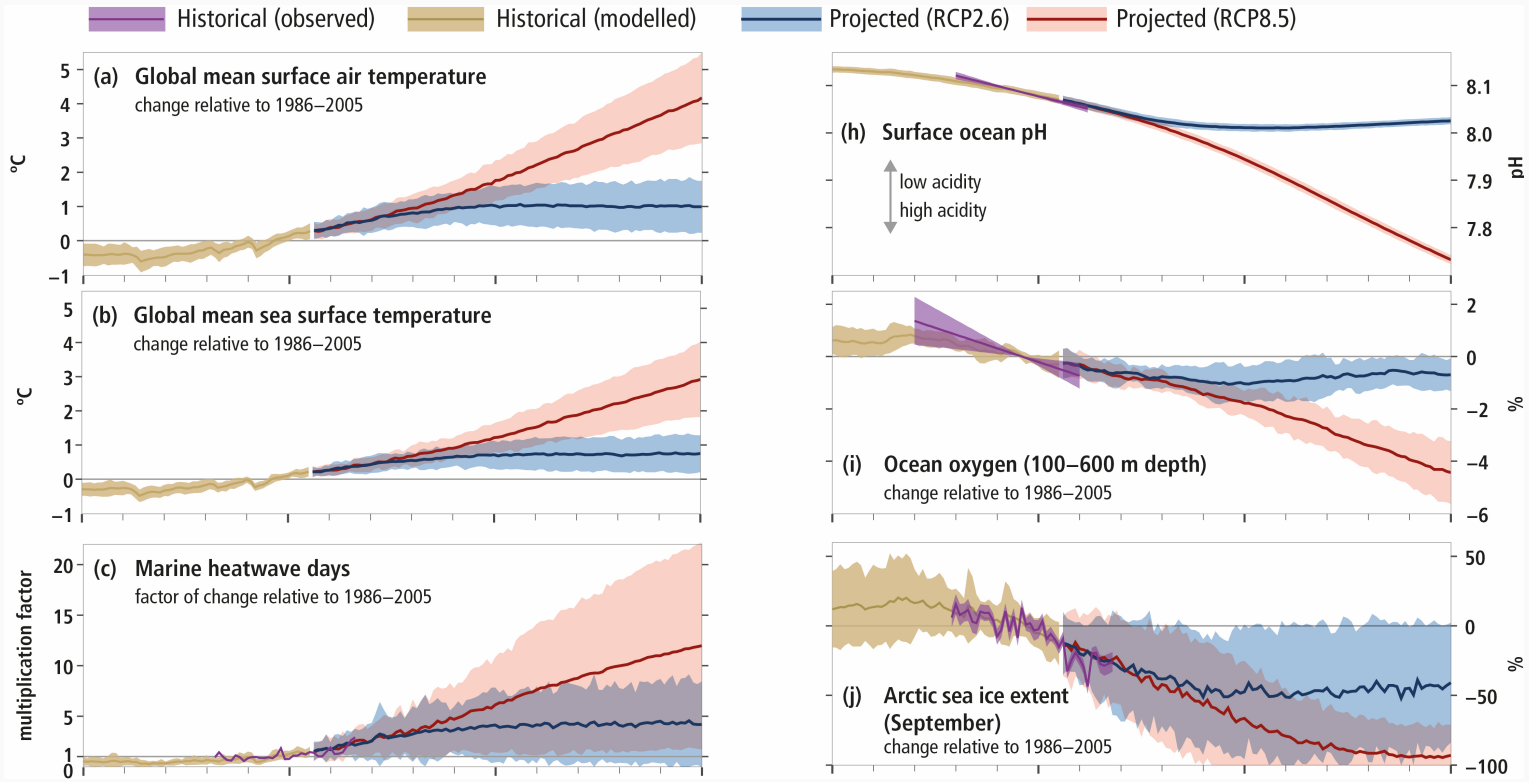


Figure 3: Ensemble uncertainty examples from the Special Report on the Ocean and Cryosphere in a Changing Climate. Bands represent 5–95% model range.

Model summaries (analyses) and inter-model variability (IMV) are used to quantify and communicate uncertainty in future climate projections.

Multi-model Ensemble Analysis

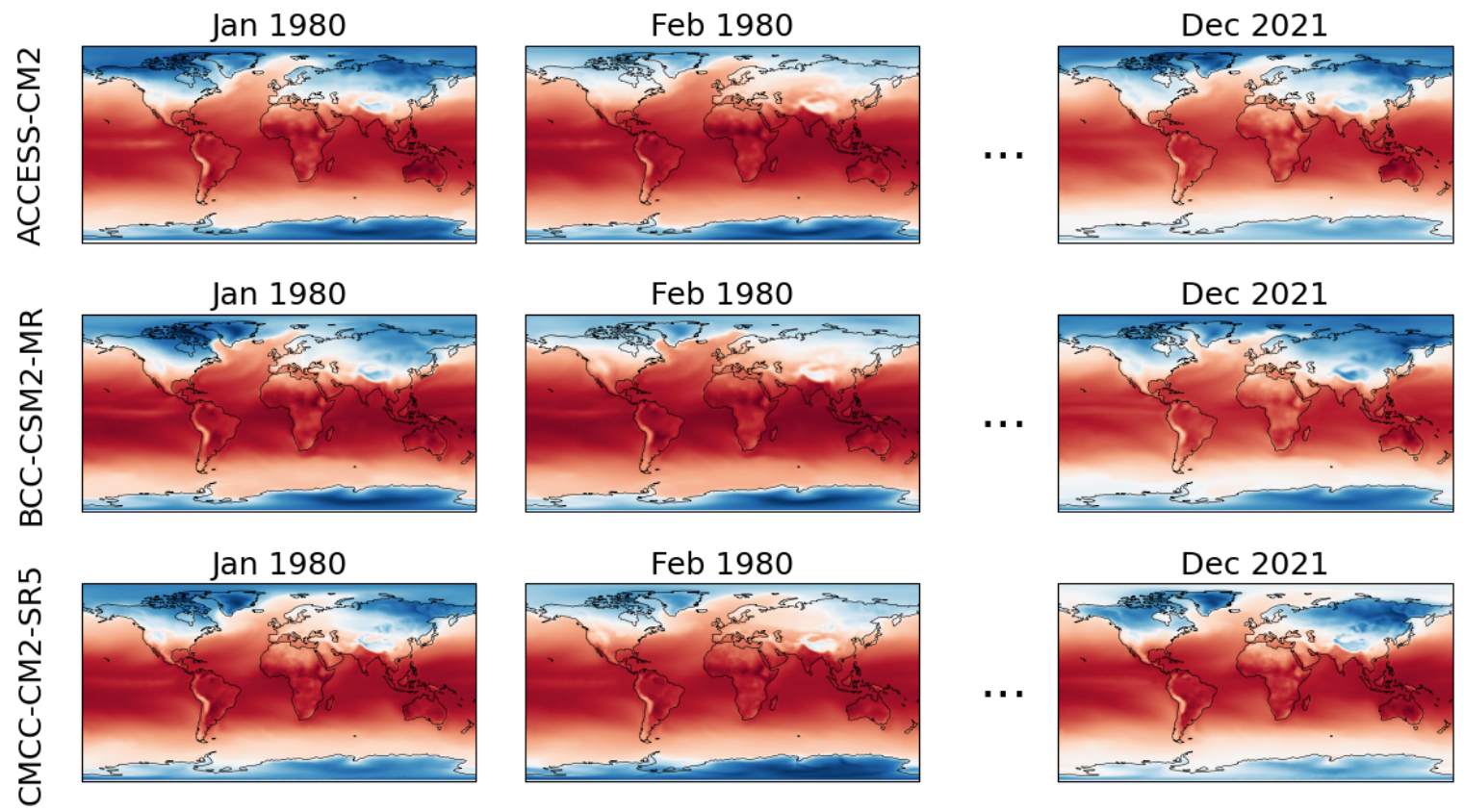


Figure 4: Climate models produce continuous spatiotemporal output. Discretized to a grid and aggregated into monthly (or daily) time steps.

Spatial fields contain far more information because they preserve local detail. Allows us to study spatially varying changes in the climate.

Problems with spatial averaging

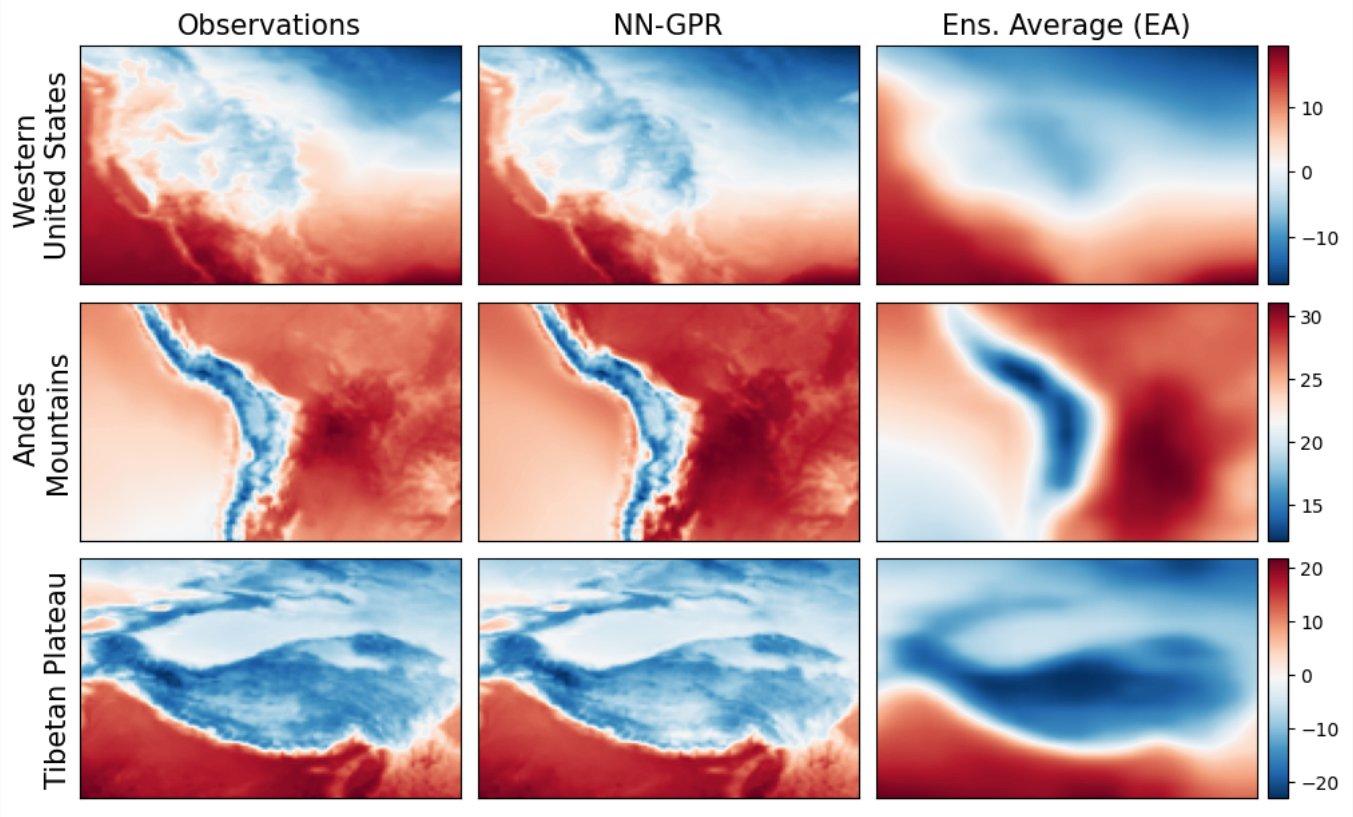


Figure 5: ERA5 reanalysis field for January 2015 (left), proposed method NN-GPR's prediction (center) and an ensemble average of the 16 climate models (right)

- Gridded observations from ERA5 reanalysis data product
- Averaging in space destroys spatial detail. Weakly represents the climate process at regional levels. Same for quantiles.

Previous work

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

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** (Rougier et al., 2013; Sansom et al., 2017; Bowman et al., 2018)
- **Regression** (Räisänen et al., 2010; Bracegirdle and Stephenson, 2012; Ghafarianzadeh and Monteleoni, 2013; Harris et al., 2023)

Regression methods have good performance (GP & CNN)

But little work on (2.). Most method have have no/poor spatial uncertainty quantification. Rely on inter-model variability (IMV)

Projection uncertainty

- IMV - A 5–95% model range represents the range of projections made by the 90% most "typical" models. But what does this really mean with regards to the underlying process?

- Statements from the IPCC AR6 such as

  > *"... lead to global warming of 3.2 [2.2-3.5]°C (5–95% range) by 2100 (medium confidence)..."*

  sound almost like a statement of confidence on the underlying process (temperature, precipitation, sea level, etc.), but they're not.

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

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

Model ensemble

- Ensemble of M climate model runs from the Coupled Modeled Intercomparison Project 6 (CMIP6)
  - Model experiments including historical simulations (1850 – near-present), and forced future projections (near-present - 2100).
  - Use monthly aggregates on 100km to 500km grids.
- Mathematically $X_{t,i}$ – 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 – an ensemble of M gridded climate model outputs observed at time t.

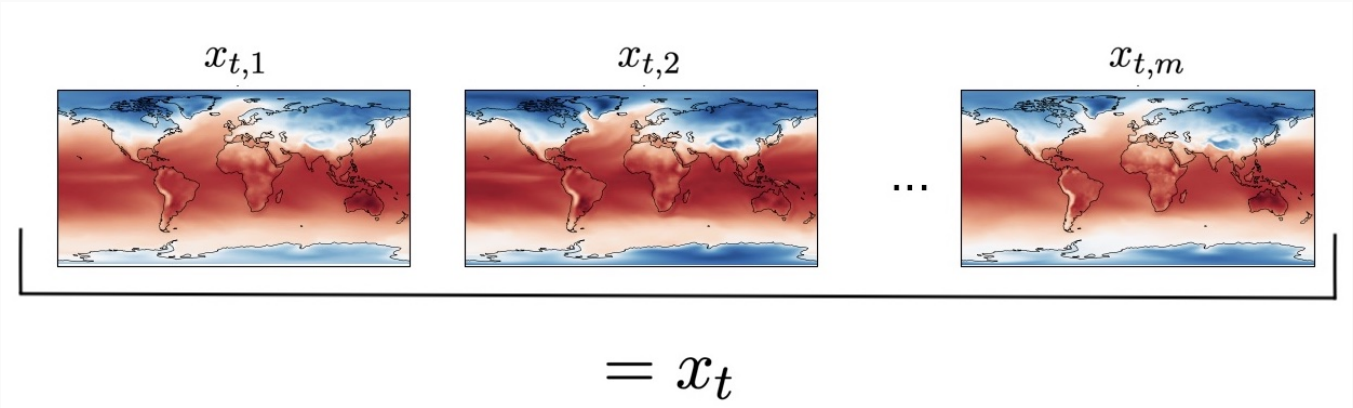


Figure 6: Example model ensemble

Response field

- As a stand in for observational data we will use reanalysis fields from ERA5 (ECMWF Reanalysis v5)
  - Reanalysis of the global climate covering January 1940 to present.
  - Hourly estimates on 31km grid resolved at 137 pressure levels up to a height of 80km.
  - Use monthly aggregates on a single pressure level
- Mathematically - Y_t – reanalysis field at time t. Each Y_t is a gridded field.

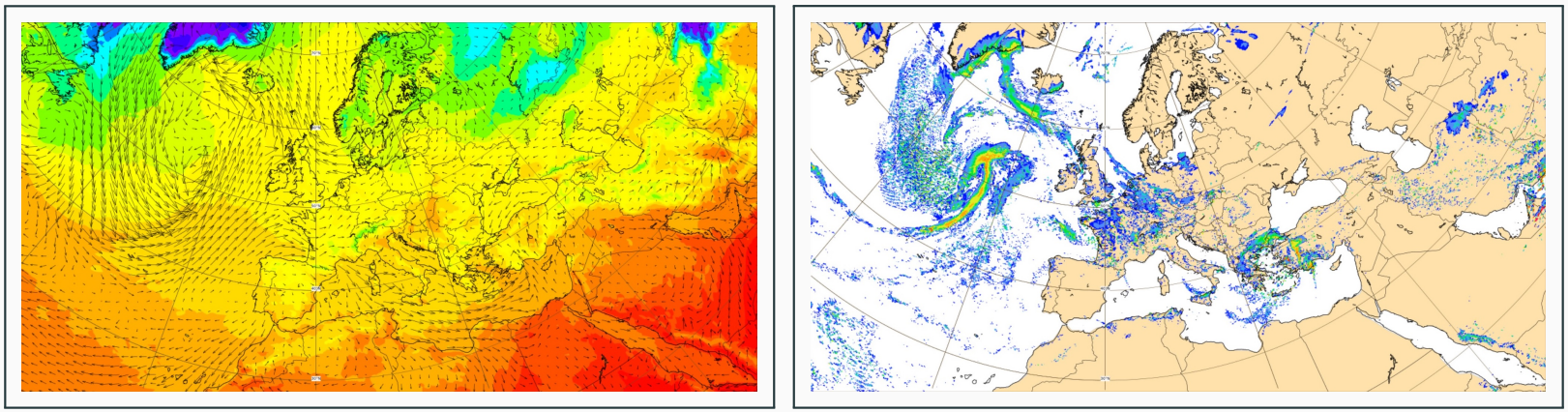


Figure 7: Left: ERA5 Surface temperature and wind speed. **Right:** ERA5 Total Precipitation

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: 1940-2020
- 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
  - No Y_t observed yet
  - Future: 2020-2100

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

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

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..."

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

- Does not require asymptotic arguements, priors, or even correctly specified models X.

Conformal inference

Multivariate conformal inference will require two things

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

1. $f_\theta(X)$ converts the ensemble X into a "prediction" of the observed Y

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

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

2. $d(Y, \hat{Y})$ identifies the $\alpha \times 100\%$ most "typical" residuals $Y - \hat{Y}$ (analogous to finding a highest density region).
   - Construct an α-level prediction set on $Y - \hat{Y}$
   - Translate to an α-level prediction set on Y

Algorithm Sketch

1. Apply f_θ to $X_1, ..., X_n$ to get predictions $Y_1, ..., \hat{Y}_n$ where

$$f_\theta(X_i) = \hat{Y}_i$$

   and (X_i, Y_i) were not used to train f_θ
2. Compute the out of sample residuals $R_1, ..., R_n$ where

$$R_i = Y_i - \hat{Y}_i$$

3. Score and order residuals from least outlying to most outlying with $d(\cdot)$. Retain the $\lceil (n+1)(1-\alpha) \rceil$ least outlying residuals

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

4. Given a new X_{n+1}, predict Y_{n+1} as

$$f_\theta(X_{n+1}) + R_{(i)} \quad \forall i \in 1, ..., \lceil (n+1)(1-\alpha) \rceil$$

   to produce an ensemble of $\lceil (n+1)(1-\alpha) \rceil$ predictions

1. Model analysis – analysis functions

- Multi-model analysis model $f_{\hat{\theta}} : X \mapsto Y$, where

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(f_\theta, D)$$

  so that $f_{\hat{\theta}}(x_t) \approx y_t$.
  - f_θ can be any* parametric or non-parametric model (model average, Gaussian process regression, convolutional neural network, etc.)
  - Only requires $\mathcal{I}(X; Y) > 0$ rather than $X \sim Y$ and does not require M large (citation to my paper)
- Caveat - To get out-of-sample predictions we will have to sample split. Might leave insufficient data for training big ML models.

$$D_{train} = \{(X_i, Y_i)\}_{i=1}^{m}$$
$$D_{cal} = \{(X_i, Y_i)\}_{i=m+1}^{n}$$

- Assume access to a "good" ensemble analysis function f_θ (Harris 2023)

2. Central regions - depth functions

Depth functions quantify the centrality of observations with respect to a reference distribution (or set of observations).

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:** $D(z + b \mid X + b) = D(z \mid X) \quad \forall b \in \mathbb{R}^d, X \sim P$
- D2 **Linear invariance:** $D(Az \mid AX) = D(z \mid X)$ for bijective linear $A : \mathbb{R}^n \mapsto \mathbb{R}^n$
- D3 **Null at infinity:** $\lim_{||z|| \to \infty} D(z \mid X) = 0$
- 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

2. Central Regions - depth functions

- Monotonicity and upper semi-continuity give us a basis for constructing prediction regions
  - Can always find $\beta \in [0, 1]$ such that $P(Y \in D_\beta(X)) \geq 1 - \alpha$ for any coverage level $\alpha \in (0, 1)$
  - $\beta = \text{Quantile}(d(z \mid P); (\lceil (n+1)(1-\alpha) \rceil)/n)$
  - Abuse notation and write $D_\alpha(X)$ for a central region with α-level coverage.

- Depth decreases monotonically from a centroid. If the error process is multi-model $\Rightarrow$ poor characterization of the distribution

- General outlier detectors (Isolation Forests (Liu et al., 2008), One Class SVMs (Chen et al., 2001), etc.(Pang et al., 2021)), allow for multi-modality in outlier scores to match the density of P.

2. Central Regions - relaxed depth functions

Relax depth to include arbitrary outlier detection by replacing condition D4 with the weaker condition

D6 **Nested level sets:** Give two upper level sets

$$D_{\alpha_1}(X) = \{z \in \mathbb{R}^d : D(z \mid X) \geq \alpha_1\}$$
$$D_{\alpha_2}(X) = \{z \in \mathbb{R}^d : D(z \mid X) \geq \alpha_2\},$$

if $\alpha_2 \leq \alpha_1$ then $D_{\alpha_1}(X) \subseteq D_{\alpha_2}(X)$.

As long as $\alpha \mapsto D_\alpha(X)$ is continuous, strictly increasing monotonic function, we can always find a central region for any confidence level for any outlier detector/centrality score/depth function.

2. Central Regions - specific relaxed depth functions

Propose a specific family of relaxed depth functions called median-max neighbor depths.

$$d(z \mid X) = 1/(1 + O(z \mid X))$$

where $O(z \mid X)$ is median ℓ_∞ distance between the query point z and nearest p neighboring points $x_{(0)}, ..., x_{(p)} \in X$.

$$O(z \mid X) = \text{Quantile}(||z - x_{(i)}||_\infty; 0.5) \quad i \in 1, ..., p$$

- $d(z \mid X)$ satisfies D1, D2 (weak), D3, D5 and D6, making it a valid relaxed depth notion.
- $d(z \mid X)$ can be adapted to the density of the residuals by varying p. Smaller = more adaptive. Typically take $p = 15$.
- Efficient at our sample sizes $n \approx 1000$, robust, adaptive to the data density, and does not require training

Algorithm – Residual Central Regions

Given: Training data $D = \{(X_i, Y_i)\}_{i=1}^{n}$, analysis function f_θ, scoring rule d, and confidence level $\alpha \in (0, 1)$

Return: Level α prediction set for the residuals $Y_i - f_\theta(X_i)$

1. Partition the training data into

$$D_{train} = \{(X_i, Y_i)\}_{i=1}^{n_1} \quad D_{cal} = \{(X_i, Y_i)\}_{i=n_1+1}^{n} \quad n_2 = n - n_1 + 1$$

2. Train the model f_θ on D_{train}
3. Compute the residual fields $R_i = Y_i - \hat{Y}_i$ on D_{cal}. Denote as R_{cal}.
4. Score and rank residuals fields using depth d to get outlier scores

$$d(R_1 \mid R_{cal}), ..., d(R_n \mid R_{cal})$$

5. Return
$\beta = \text{Quantile}(d(R_1 \mid R_{cal}), ..., d(R_n \mid R_{cal}), (\lceil (n_2 + 1)(1 - \alpha) \rceil)/n_2)$

Conformal Ensembles

From here we can construct two objects: the conformal prediction region and the conformal ensemble

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

$$D_\alpha(X) = \{z : d(z \mid R_{cal}) \geq \beta\}$$

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

$$C_\alpha(X) = \{z + \hat{Y} : d(z + \hat{Y} \mid R_i + \hat{Y}, ..., R_{n-m} + \hat{Y}) \geq \beta\}$$

is a valid prediction region on Y.

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

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

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

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

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 α level coverage on $Y \in D_{test}$.

Exchangeability issues

- What if the residuals are not exchangeable?

- Can happen if our data experiences **distribution shift**

$$P_{tr}(X, Y) \neq P_{te}(X, Y)$$

  i.e the joint distribution of the predictors X and targets Y is different in the train and test sets.
  - **Always present due to climate change**
  - Simulated in the models (X), realized in the observations (Y)

- If a model is not robust to distribution shift, then the residuals will not be exchangeable and coverage is not guaranteed.

Exchangeability issues

- Conformal methods are sensitive to distribution shift. Can mitigate these issues by either
  1. Widening the prediction interval (Current approach)
  2. Developing models that are robust to covariate shift (ongoing work)

- Theoretically – Inflate the α level based on the distance between (X_{test}, Y_{test}) and the calibration data (X_{cal}, Y_{cal}) (Angelopoulos and Bates, 2021)

$$\alpha \mapsto \alpha + 2 \sum_{i=1}^{n_2} w_i \epsilon_i$$

  where $\epsilon_i = d_{TV}((X_{test}, Y_{test}), (X_{cal}, Y_{cal}))$ and $w_1 + w_2 + ... = 1$.

- Caveat – ϵ_i can't be computed (requires knowledge of Y_{test}). Can approx. in time series, doesn't work for long range forecasting.

Exchangeability issues

- Special case – If the distribution shift is purely in X, i.e. $P(X_{cal}) \neq P(X_{test})$ then we have *covariate shift*.

- Computable approach to correcting for covariate shift (Tibshirani et al., 2019). Re-weight the conformal scores by

$$w(x) = \frac{p_{test}(x)}{p_{cal}(x)} \propto \frac{P(C = 1 \mid X = x)}{P(C = 0 \mid X = x)}$$

  where $C = 1$ indicates membership in the test set.

- Train a classifier to distinguish between test and calibration data and uses the estimated probabilities.

- In practice... unable to train an accurate, well calibrated classifier. Not much empirical coverage gain.

Proposal: Reweight scores with more depths

- **Proposal:** Re-weight the residual depths by a covariate depth ratio (inspired by previous approach)
- We assume that

$$\frac{d(X \mid X_{test})}{d(X \mid X_{cal})} \approx \frac{d(r \mid R_{test})}{d(r \mid R_{cal})},$$

  i.e. the divergence between covariates (cal to test) is proportional the divergence between residuals.
- Re-weight the calibration depths to "look like" test depths

$$d^*(r \mid R_{cal}) = d(r \mid R_{cal}) \frac{d(X \mid X_{test})}{d(X \mid X_{cal})} \approx d(r \mid R_{test})$$

- Empirical testing shows the assumed relationship is relatively accurate, but the adjusted depths tend to over-cover. No extra training though!

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)
  - Repeat for all models to get jackknife-like estimates
- Compare our conformal ensemble method (Conf.) against the inter-model variability (IMV) given three different analyses algorithms (Harris et al., 2023)
  - Ensemble Average - pointwise average
  - Neural Network Gaussian Process - NN-GPR
  - Deep Convolutional Neural Network - CNN
- Apply on mean 2-meter air surface temperature (tas), maximum 2-meter air surface temperature (tasmax), and Total Precipitation (pr)

Numerical experiments

- Train on historical period (1950-2021) match reanalysis data availability
  - Monthly model means. 852 historical observations. Use the last 200 as a calibration dataset and the first 652 as a proper training set.
- Test on future simulations (2021-2100) based on SSP245
  - SSP245 – Shared Socioeconomic Pathway 2 with Representative Concentration Pathway (RCP) 4.5 (medium plausible scenario)
- All models regridded to $80x120$ grid cells (pixels). Done for computational reasons, evidence this improves generalization.
- Use the proposed median-max nearest neighbor depth with 15 neighbors. Unless noted, construct interval with $\alpha = 0.1$.

UQ metrics

Five metrics for measuring different qualities of the uncertainty quantification (Gneiting and Katzfuss, 2014)

1. **CRPS** - Continuous Ranked Probability Score. Measures how close empirical CDF is to true CDF.
2. **PIT** - Distance between the probability integral transform and a uniform distribution. Lower PIT scores means the forecast is more calibrated.
3. **Dispersion** - Difference between PIT variance and uniform variance (1/12). Negative means underdispersed, Positive means overdispersed.
4. **Sharpness** - Width of the prediction interval. Measures precision of UQ. Sharper is better, provided the forecast is calibrated.
5. **Coverage** - 90% coverage of prediction intervals (only for Conf.)

Coverage – white noise

- White noise experiment to verify depth based UQ on spatial ensembles

$$X_{cal} \sim N(0_{1000\times30\times10\times10}, 1)$$
$$Y_{cal} = 0_{1000\times10\times10}$$
$$X_{test} \sim N(0_{1000\times30\times10\times10}, 1)$$
$$Y_{test} = 0_{1000\times10\times10}$$

- $\hat{Y} = \sum_{i=1}^{30} X[,i,,]$
- Repeat 500 times

(almost) perfect out of sample coverage $\Rightarrow$

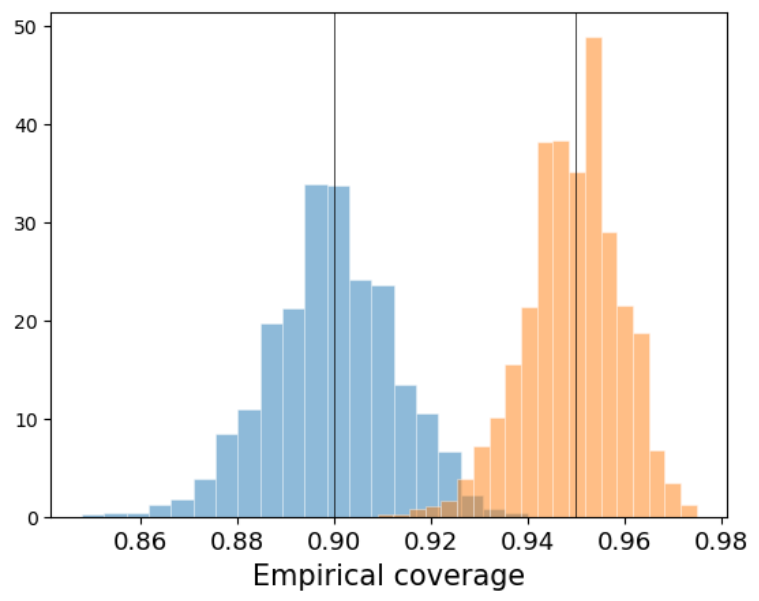


Figure 8: Empirical coverage on pure white noise and a "model average" predictor on 10x10 spatial fields.

Theoretical	Empirical
$(1-\alpha) = 0.90$	0.90000004
$(1-\alpha) = 0.95$	0.95000005

Coverage – Models

$(1-\alpha) = 0.9$	tas	tasmax	pr
Ens. Conf.	0.882	0.886	0.886
NN-GPR Conf.	0.867	0.861	0.896
CNN Conf.	0.897	0.883	0.887
(Adj) Ens. Conf.	0.965	0.953	0.919
(Adj) NN-GPR Conf.	0.953	0.941	0.915
(Adj) CNN Conf.	0.957	0.953	0.921

Table 1: Coverage of Conf. regions for three models (Ens. Avg, NN-GPR, CNN) on three climatic variables (tas, tasmax, pr) averaged over (30, 19, 31) model experiments respectively

- Consistent empirical coverage across model and data type
- Slight undercoverage due to covariate shift
- Adjustment overcorrects undercoverage

Metrics on TAS

	RMSE	CRPS	Sharp.	PIT	Disp.
Ens. IMV	1.0619	0.5587	4.0719	0.0660	-0.0153
Ens. Conf.	1.0619	0.3992	3.2209	0.0169	0.0009
GP IMV	0.6540	0.5807	4.0719	0.0710	-0.0150
GP Conf.	0.6540	0.3456	2.8462	0.0197	0.0037
CNN IMV	0.7016	0.5919	4.0719	0.0510	-0.0053
CNN Conf.	0.7016	0.3689	3.1571	0.0136	0.0004

Table 2: UQ metrics for Ens. Avg (Ens.), NN-GPR, and a deep CNN (CNN) using centered IMV and a Conformal ensemble (Conf.) to quantify uncertainty in tas. Metrics averaged over entire prediction interval (2020-2100).

- Conf. Ensemble results in lower CRPS, PIT, and dispersion with sharper regions compared to IMV for all models and variables.
- Rewards more accurate models (lower RMS $\Rightarrow$ lower CRPS)

Metrics on TASMAX

	RMSE	CRPS	Sharp.	PIT	Disp.
Ens. IMV	1.0864	0.5959	4.0057	0.0701	-0.0128
Ens. Conf.	1.0864	0.3743	2.9839	0.0149	0.0033
GP IMV	0.6382	0.6532	4.0057	0.0867	-0.0072
GP Conf.	0.6382	0.3229	2.6294	0.0190	0.0059
CNN IMV	0.6821	0.5945	4.0057	0.0550	-0.0079
CNN Conf.	0.6821	0.3429	2.9488	0.0114	0.0018

Table 3: UQ metrics for Ens. Avg (Ens.), NN-GPR, and a deep CNN (CNN) using centered IMV and a Conformal ensemble (Conf.) to quantify uncertainty in tasmax. Metrics averaged over entire prediction interval (2020-2100).

- Conf. Ensemble results in lower CRPS, PIT, and dispersion with sharper regions compared to IMV for all models and variables.
- Rewards more accurate models (lower RMS $\Rightarrow$ lower CRPS)

Metrics on PR

	RMSE	CRPS	Sharp.	PIT	Disp.
Ens. IMV	1.3747	0.6930	4.6913	0.0513	0.0141
Ens. Conf.	1.3747	0.5087	5.2226	0.0028	-0.0018
GP IMV	1.2261	0.6305	4.6913	0.0392	-0.0021
GP Conf.	1.2261	0.4871	5.1665	0.0048	-0.0058
CNN IMV	1.2924	0.6304	4.6913	0.0268	0.0027
CNN Conf.	1.2924	8.8570*	28.1426*	0.1646	0.0118

Table 4: UQ metrics for Ens. Avg (Ens.), NN-GPR, and a deep CNN (CNN) using centered IMV and a Conformal ensemble (Conf.) to quantify uncertainty in pr. Metrics averaged over entire prediction interval (2020-2100).

- Conf. Ensemble results in lower CRPS, PIT, and dispersion with sharper regions compared to IMV for all models and variables.
- Rewards more accurate models (lower RMS $\Rightarrow$ lower CRPS)

Metrics over time (tas)

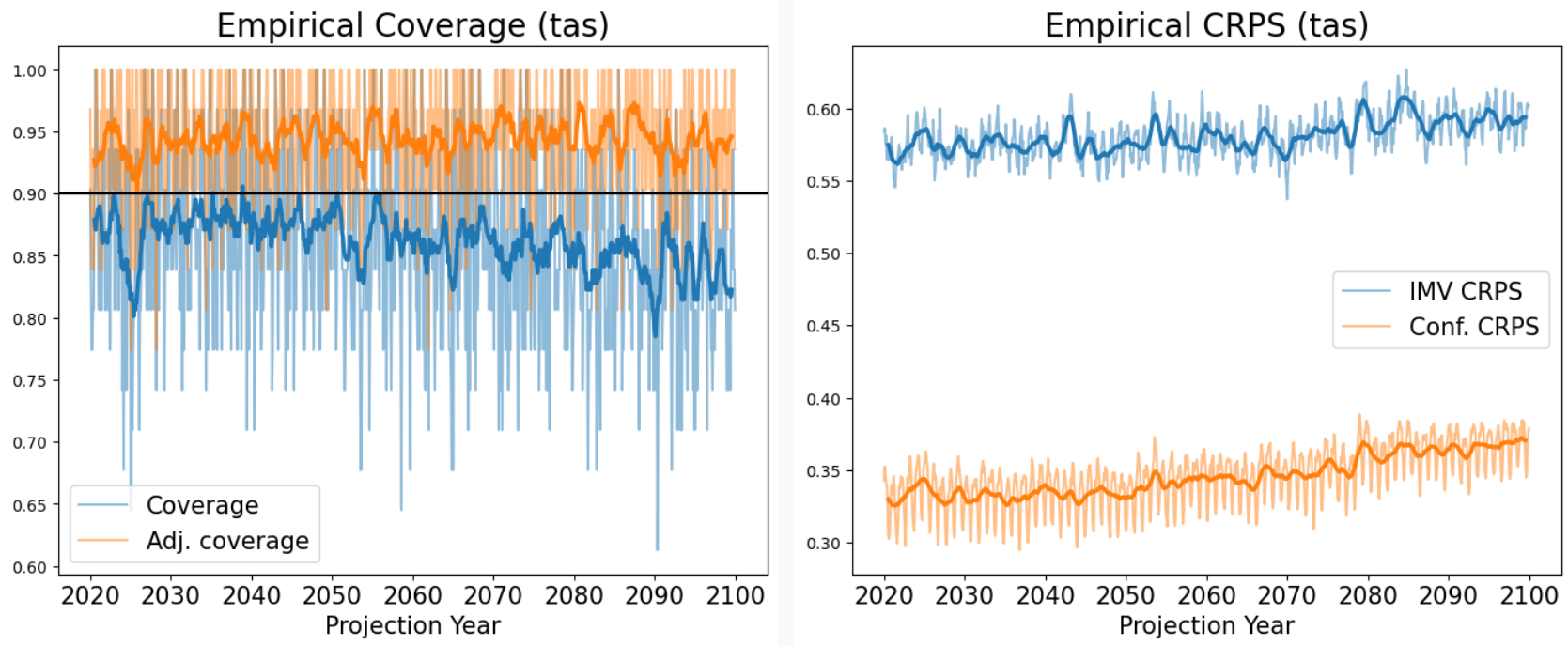


Figure 9: left: Coverage over time or an NN-GPR base using a Conf. Ensemble with raw score quantiles (blue) and adjusted score quantiles (orange). **Right:** CRPS over time for an NN-GPR base using IMV (blue) and a Conf. Ensemble (orange).

Conformal ensembles have consistently lower CRPS scores across all models, time points, and variables compared with IMV.

Metrics over time (tasmax)

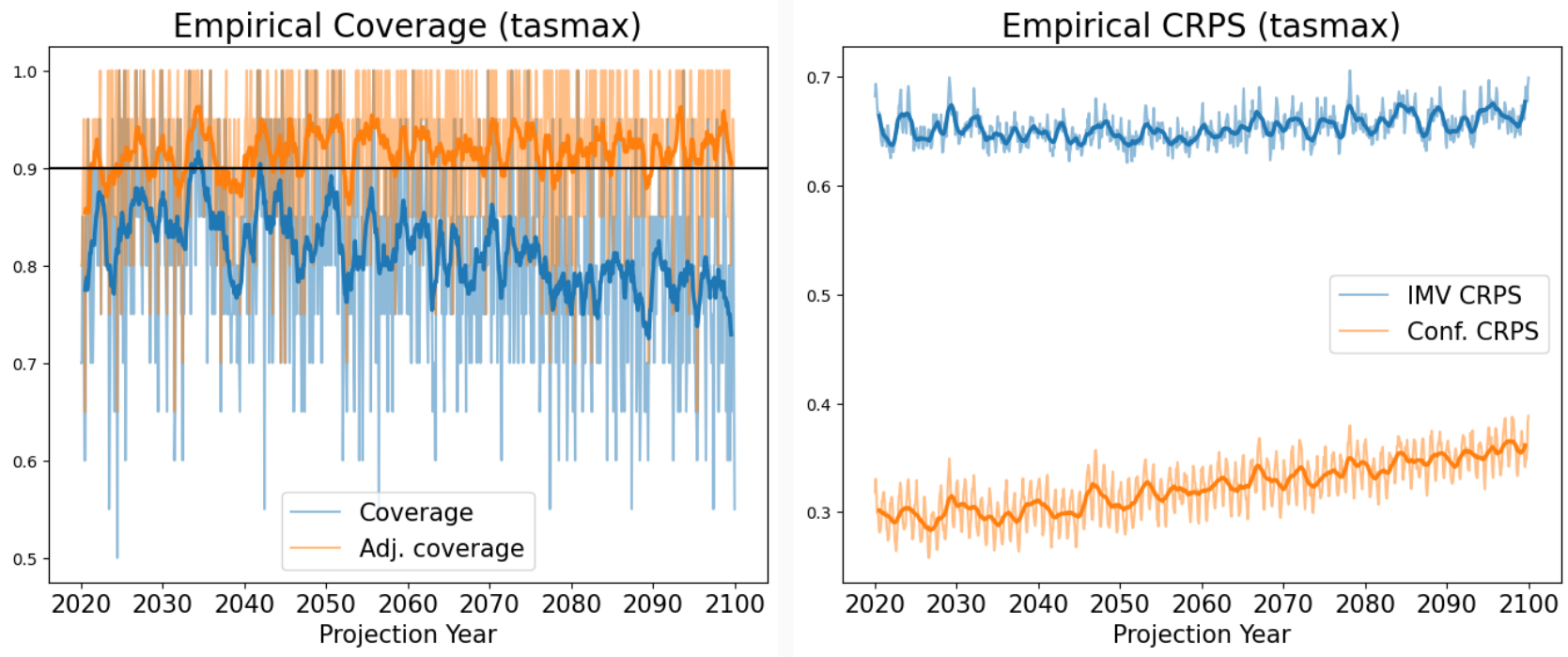
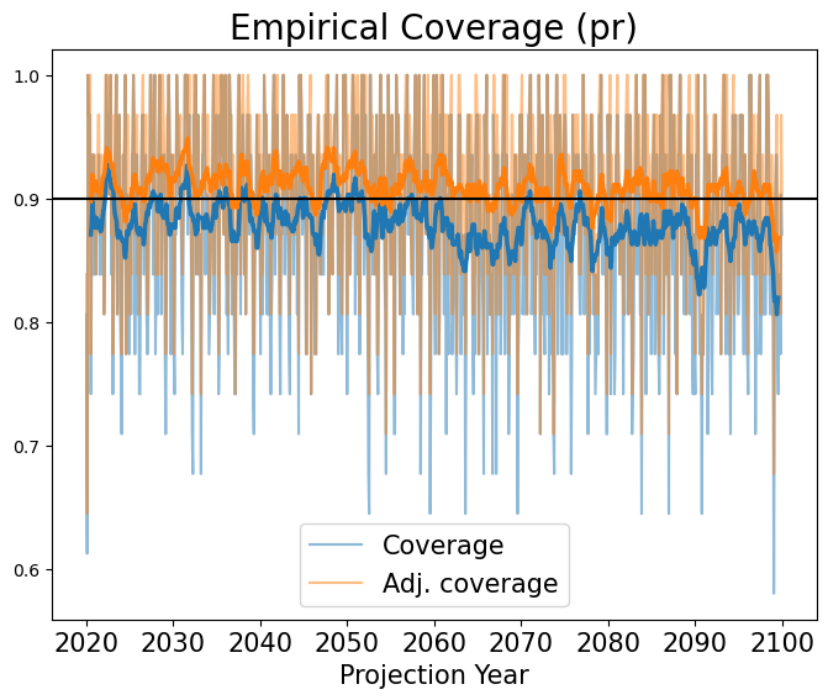


Figure 10: left: Coverage over time or an NN-GPR base using a Conf. Ensemble with raw score quantiles (blue) and adjusted score quantiles (orange). **Right:** CRPS over time for an NN-GPR base using IMV (blue) and a Conf. Ensemble (orange).

Conformal ensembles have consistently lower CRPS scores across all models, time points, and variables compared with IMV.

Metrics over time (pr)



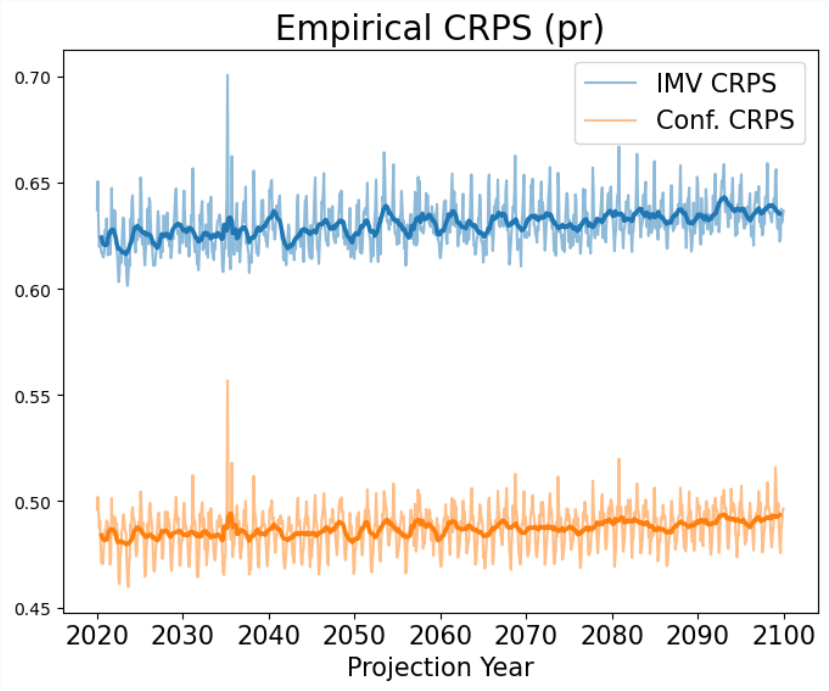


Figure 11: left: Coverage over time or an NN-GPR base using a Conf. Ensemble with raw score quantiles (blue) and adjusted score quantiles (orange). **Right:** CRPS over time for an NN-GPR base using IMV (blue) and a Conf. Ensemble (orange).

Conformal ensembles have consistently lower CRPS scores across all models, time points, and variables compared with IMV.

Metrics over space (CRPS)

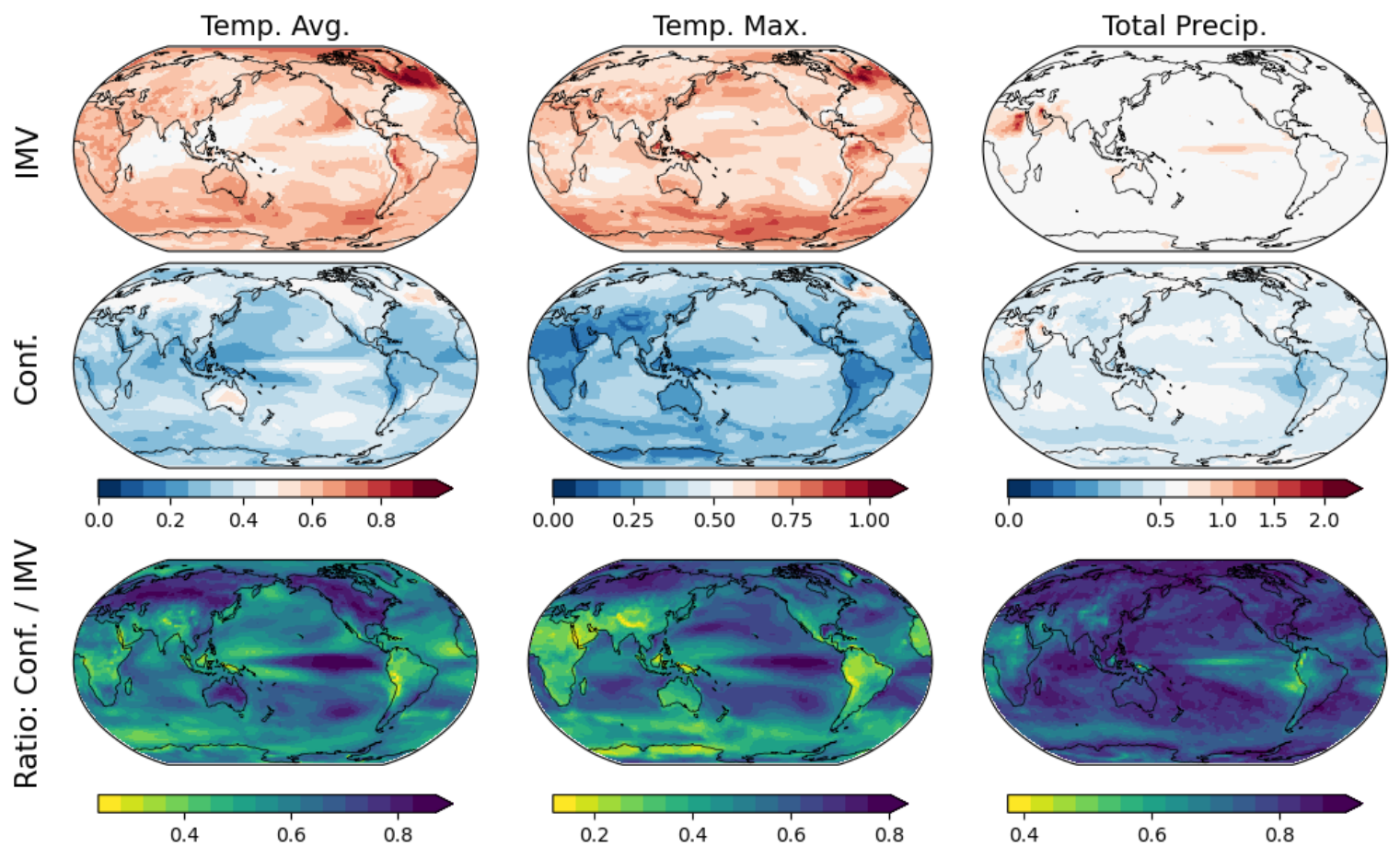


Figure 12: Top: IMV based CRPS scores using an NNGP model averaged over time (2020-2100) for all spatial locations. **Middle:** Conformal based CRPS scores using an NNPG model averaged over time(2020-2100) for all spatial locations. **Bottom:** Ratio Conf. / IMV of the two fields.

Metrics over space (Sharpness)

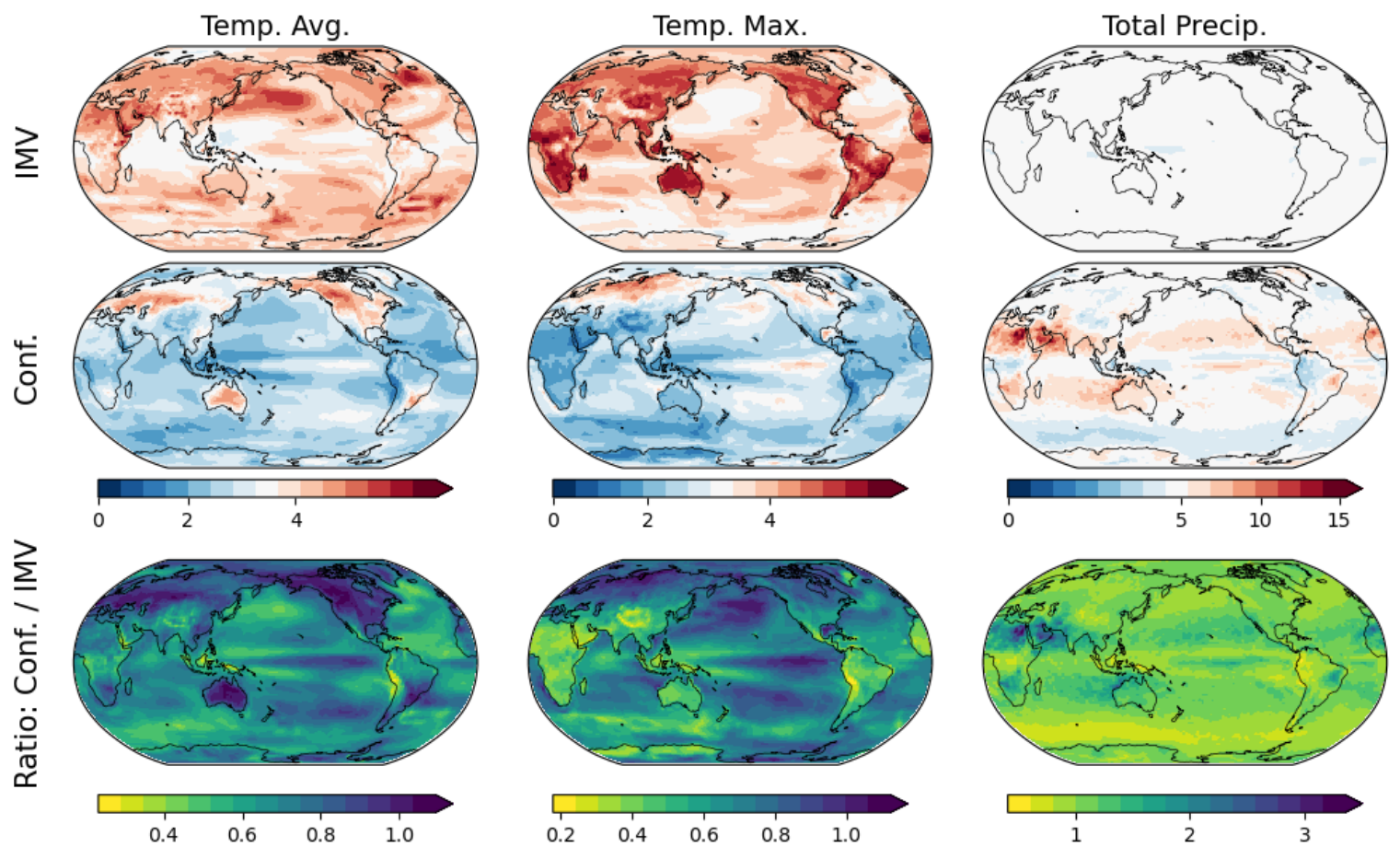


Figure 13: Top: IMV based Sharpness scores using an NNGP model averaged over time (2020-2100) for all spatial locations. **Middle:** Conformal based Sharpness scores using an NNPG model averaged over time(2020-2100) for all spatial locations. **Bottom:** Ratio Conf. / IMV of the two fields.

Metrics over space (PIT)

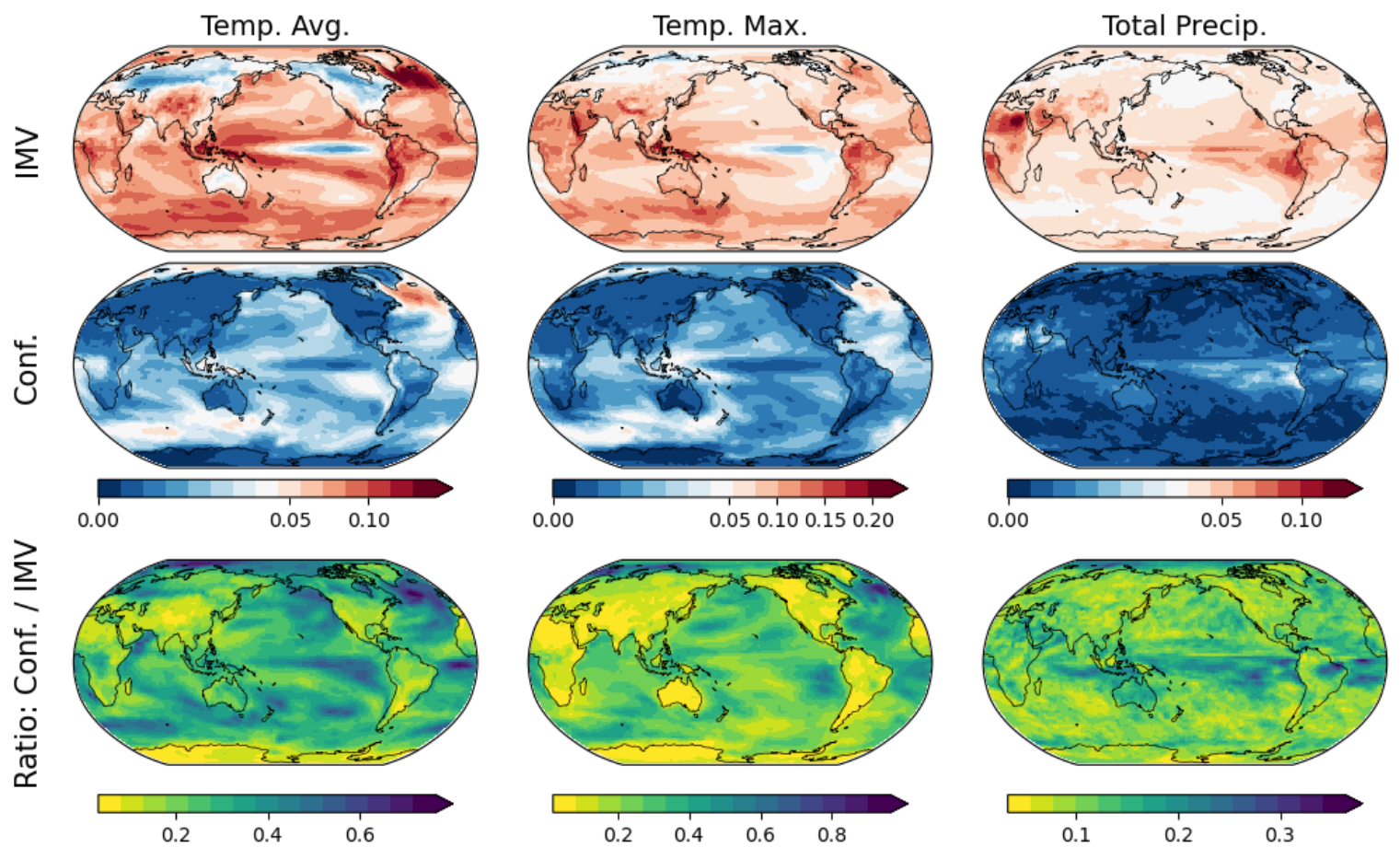


Figure 14: Top: IMV based PIT scores using an NNGP model averaged over time (2020-2100) for all spatial locations. **Middle:** Conformal based PIT scores using an NNPG model averaged over time(2020-2100) for all spatial locations. **Bottom:** Ratio Conf. / IMV of the two fields.

Application - Global multi-model analysis

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

- Use NN-GPR to make future projections under SSP245.
  - Project 2-meter surface temperature (TAS) and Total Precipitation (PR) monthly averages
  - Take the same 31 and 32 member model ensemble from the experiments as input
  - Predict the corresponding ERA5 reanalysis fields.
  - ERA5 much higher resolution than any of the models (1440×720 grid points).

- Compare global projections with Ens. Avg. + IMV
  - Train on historical period (1940-1995)
  - Calibrate Conf. Ens. on (1995-2021)

Global multi-model analysis

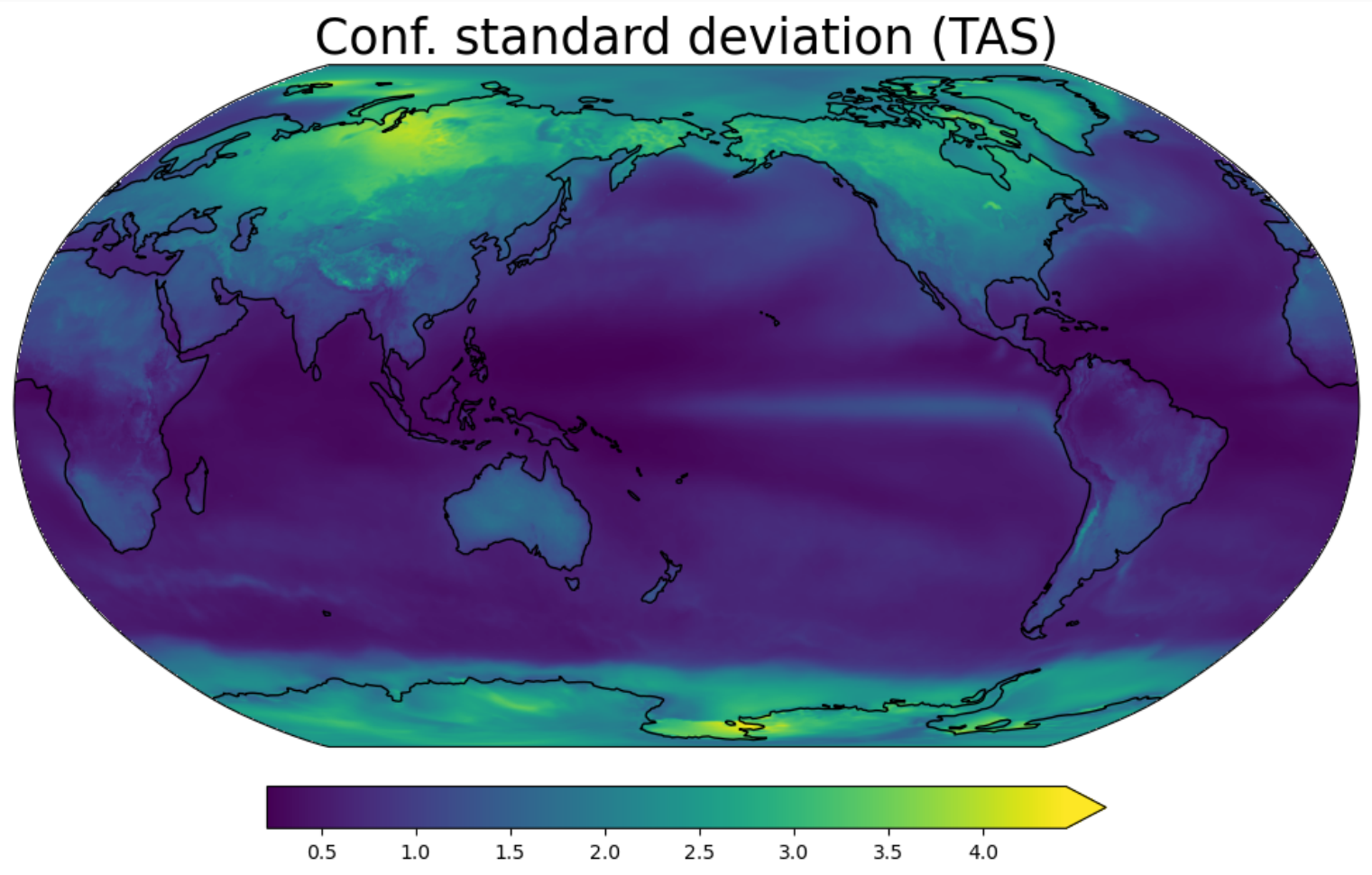


Figure 15: Pointwise standard deviation of the tas conformal ensemble using NN-GPR as the analysis function.

Global multi-model analysis

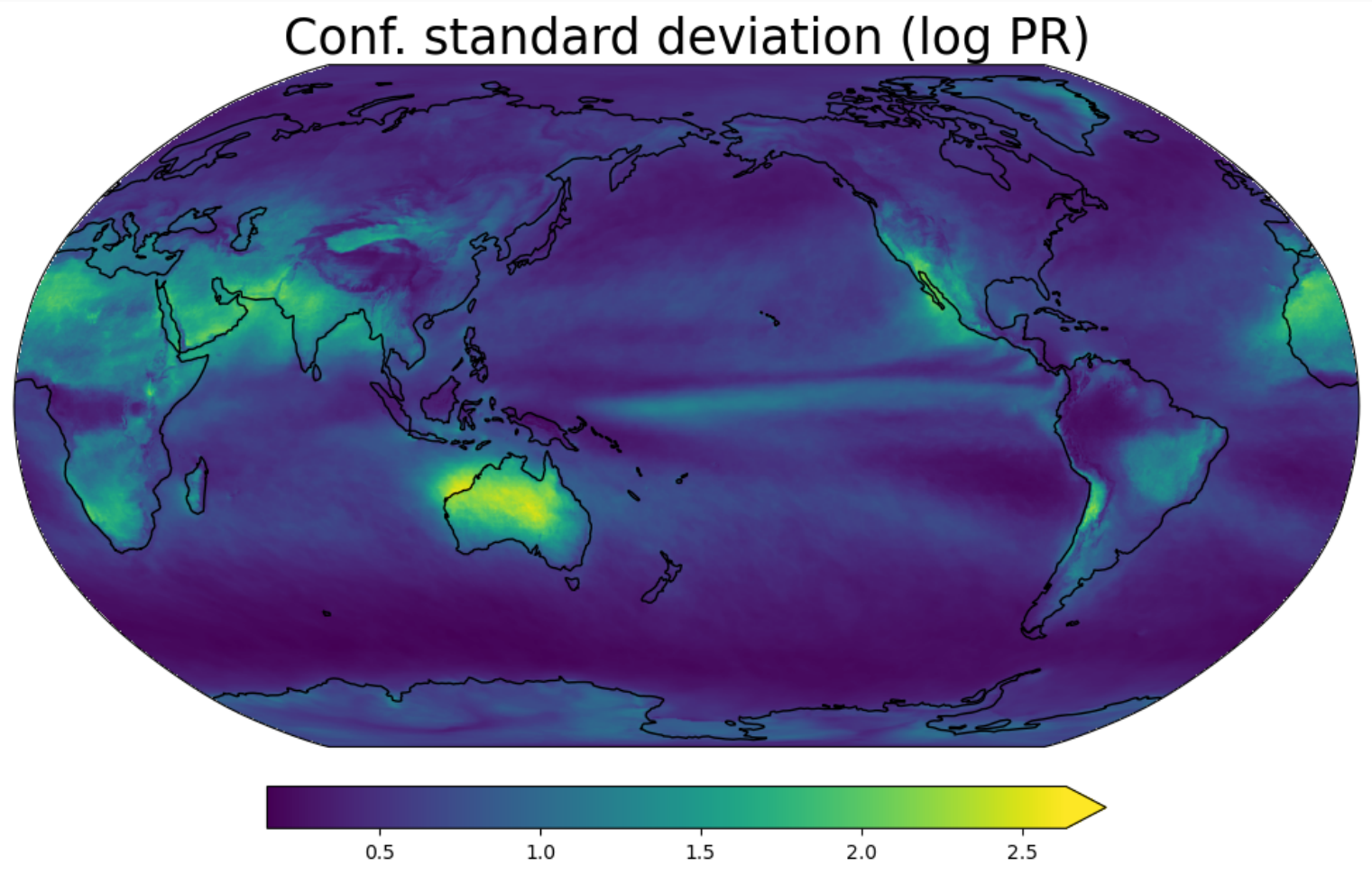


Figure 15: Pointwise standard deviation of the pr conformal ensemble using NN-GPR as the analysis function.

Global multi-model analysis

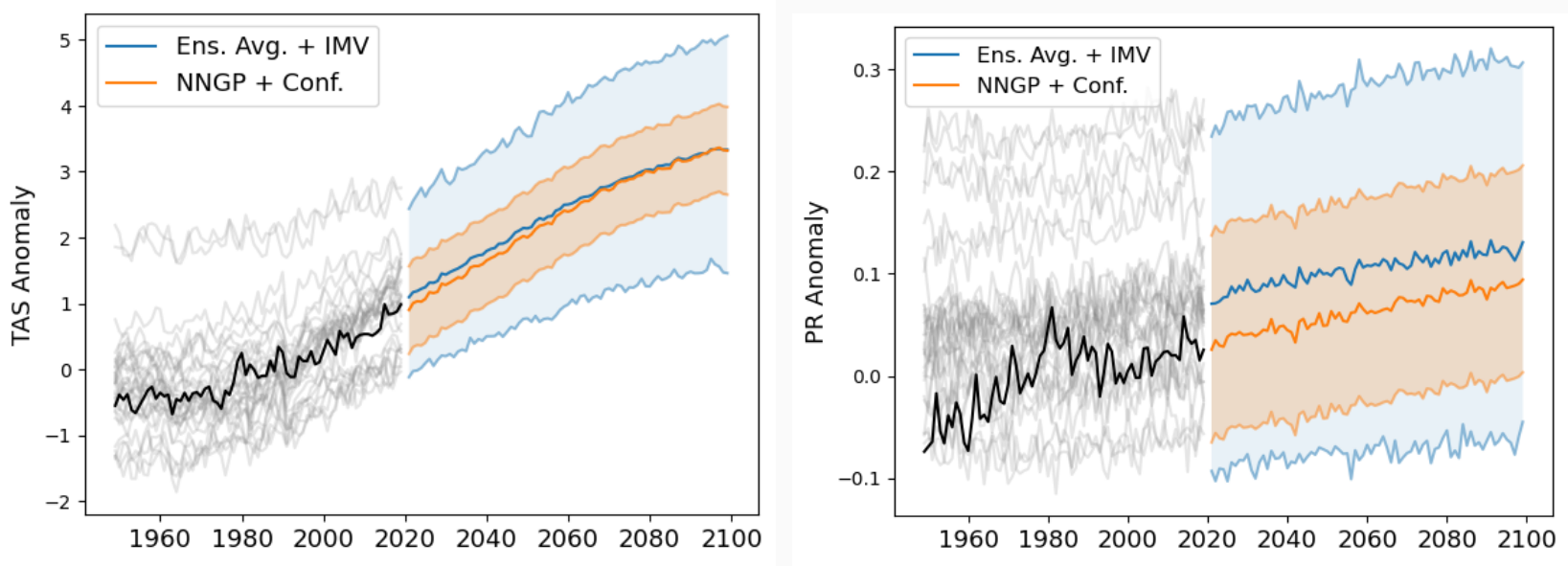


Figure 15: Left: Pointwise standard deviation of the tas conformal ensemble using NN-GPR as the analysis function. **Right:** Same but for log(pr). Note: one model excluded from pr plot due to extreme outlyingness.

- NNGP predictions closely track model average (good), but have significantly lower predictive variance.

Summary

- Multi-model ensemble analysis is critical for optimally projecting climate and for quantify uncertainty in projections
- Standard analysis technique (Ens. Avg. + IMV) is insufficient for spatiotemporal projection. Model based techniques improve but have poor spatial UQ.
- We introduce a conformal inference based approach to UQ called conformal ensembles that works for any model analysis method (ensemble averaging, regression, GPR, NNs, etc.)
  - Relax depth to include more general outlier detectors
  - Adjustment technique to widen prediction intervals to guard against distribution shift
- Projection intervals have frequentist coverage guarantees (assuming exchangeability)

Summary

- 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
  - Conformal ensembles improves over IMV across all variables, models, time points, and spatial locations. Lower CRPS, Lower PIT, Correct sharpness.
  - (not shown) Conf. CRPS decrease with model ensemble size.
  - (not shown) Larger calibration sets slightly decrease CRPS and improve coverage, until too large and not enough obs for training.
- Application to reanalysis data shows
  - Comparable projections to the ensemble mean. Good since Integration method should probably not change the overall global projection too much. Regional variations are fine.
  - Higher precision using a more "accurate" projection method. Good if coverage guarantees hold.

NNGPR reference

Harris, T., Li, B., & Sriver, R. (2023). *Multimodel ensemble analysis with neural network Gaussian processes*. Annals of Applied Statistics, 17(4), 3403-3425.