A statistical learning approach to multi-model ensemble analysis

Trevor Harris

October 1, 2024

University of Connecticut
Statistics Department

ENVR, Boulder, CO

Combining models with uncertainty

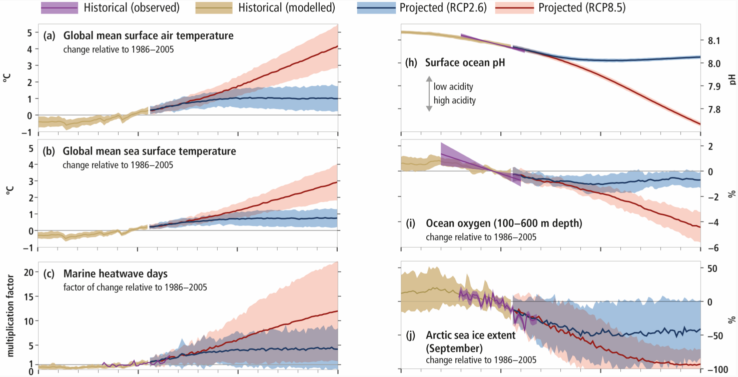
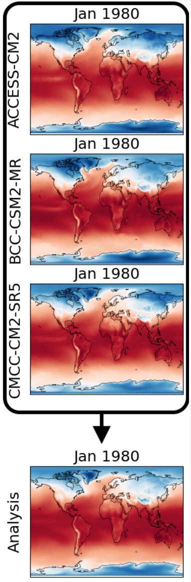


Figure 1: CMIP6 projections for an array of climate variables. Large ensembles are condensed down into a single projection (bold line) with uncertainty (shading).

Ensemble analysis is a two part problem

1. Distill GCMs down to a single projection
2. Quantify uncertainty around the projection

A statistical learning approach



Treat ensemble analysis as a **prediction** problem.

- **Data**: $D_{train} = \{(X_t, Y_t)\}_{t=1}^{n}$
  - X_t – Ensemble of GCM output (fields)
  - Y_t – Reanalysis / Observations (field)
- **Goal**: Learn f_θ s.t. $\ell(f_\theta(X_t), Y_t)$ is minimized
  - f_θ is our ensemble analysis function
  - Part 1. f_θ maps GCM ensemble to single field conditioned on observations
  - Part 2. Prediction intervals / sets of f_θ quantify uncertainty
- **Assumptions**:
  - Learnability: $h(Y \mid X) < h(Y)$ - Information from X reduces uncertainty about Y
  - Stationarity: f_θ generalizes to $D_{test} \sim D_{train}$

Part 1. Optimally combining models

What sort of analysis function f_θ is best?

- Can be nearly any prediction algorithm (weighted mean, GLM, GPR, CNN, etc.). Deterministic or stochastic.
  - Required: Take a GCM ensemble X_t and return a field $f_\theta(X_t) = \hat{Y}_t$
  - Prefer: f_θ to be as accurate as possible (sharper prediction sets), robust to covariate shift (intrinsic to climate data), robust to the curse of dimensionality (high dimensional regression), trainable from few observations.
- Proposal: Gaussian Process regression (GPR) with a neural network derived kernel (NNGP)
  - Posterior predictive distribution exactly equal to the posterior predictive distribution of an infinitely wide BNN
  - Asymptotic approximation to a fully trained, finite width NN
  - Can be "trained" to learn $f(X_t) \approx Y_t$ with very little data, inbuilt UQ, robust-ish.

Statistical Model - NN-GPR

Let $y_t(s)$ denote location $s \in s_1, ..., s_d$ in the reanalysis field y_t. We model each $y_t(s)$ as

$$\begin{aligned} y_t(s) &= \bar{\mathbf{x}}_t \boldsymbol{\beta} + f_s(x_t) + \epsilon_{s,t}, \\ f_s &\sim \mathcal{GP}\left(0, \mathbf{K}_\phi\right), \\ \epsilon_{s,t} &\sim \text{ iid } N(0, \sigma^2) \end{aligned} \tag{1}$$

- $\bar{\mathbf{x}}_t$ is the length-m vector of the climate model means at time t, i.e., $\bar{\mathbf{x}}_t = (\bar{x}_{t,1}, ..., \bar{x}_{t,m})^T$ for which $\bar{x}_{t,i}, 1 \leq i \leq m$ represents the spatial mean of the ith climate model output at time t
- $f_s \sim \mathcal{GP}\left(\mathbf{0}, \mathbf{K}_\phi\right)$ is the Gaussian process prior corresponding to an infinitely wide Bayesian neural network with a pre-specified architecture.
- $\mathbf{K}_\phi$ shared across all locations $s \in s_1, ..., s_d$ (same ϕ at all locations).

Point Estimation - T2M

(↓) Mean Squared Error (MSE) - T2M

Model	2030	2040	2050	2060	2070	2080	2090	2100
NN-GPR	**1.91**	**1.97**	**2.10**	**2.27**	**2.37**	**2.53**	2.68	2.84
LM	2.29	2.28	2.38	2.51	2.54	2.57	**2.62**	**2.71**
WEA	3.29	3.27	3.40	3.54	3.54	3.60	3.62	3.67
EA	5.98	5.87	5.96	6.04	6.00	6.03	5.97	5.99
GPSE	**1.91**	**2.01**	2.26	2.57	2.85	3.23	3.60	3.96
GPEX	**1.89**	**1.97**	2.19	2.44	2.65	2.90	3.16	3.40
CNN	2.78	2.75	2.79	2.95	2.94	2.97	3.01	3.08
DELT	3.07	3.05	3.17	3.31	3.30	3.36	3.40	3.46

Table 1: Point estimation metrics, averaged over perfect model experiments, broken out by decade. NN-GPR is our method, LM is pointwise linear regression, WEA is a reliability weighted ensemble average, and EA is the naive ensemble average. Arrows indicate whether higher (↑) or lower (↓) numbers are better.

Point Estimation - PR

(↓) Mean Squared Error (MSE) - PR

Model	2030	2040	2050	2060	2070	2080	2090	2100
NN-GPR	**3.84**	**3.97**	**4.05**	**4.28**	**4.47**	**4.59**	**4.68**	**4.83**
LM	4.41	4.58	4.64	4.84	4.99	5.08	5.16	5.29
WEA	4.97	5.14	5.22	5.43	5.58	5.65	5.76	5.85
EA	5.84	6.03	6.13	6.33	6.48	6.57	6.70	6.76
GPSE	**3.88**	**4.02**	**4.08**	**4.33**	**4.52**	**4.63**	**4.73**	**4.87**
GPEX	**3.86**	**3.99**	**4.06**	**4.31**	**4.49**	**4.61**	**4.71**	**4.86**
CNN	4.70	4.87	4.92	5.15	5.34	5.41	5.49	5.63
DELT	5.15	5.31	5.40	5.60	5.74	5.85	5.97	6.05

Table 2: Point estimation metrics, averaged over perfect model experiments, broken out by decade. NN-GPR is our method, LM is pointwise linear regression, WEA is a reliability weighted ensemble average, and EA is the naive ensemble average. Arrows indicate whether higher (↑) or lower (↓) numbers are better.

Global multi-model analysis

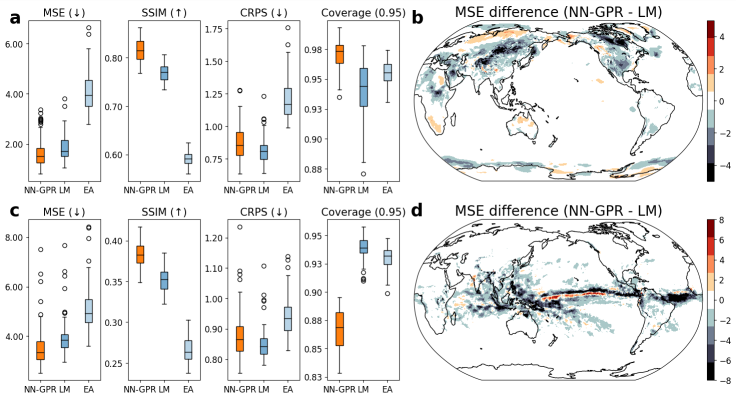


Figure 2: Boxplots for T2M (a) and PR (c) comparing the accuracy and UQ measures for each method on the reanalysis test data (2015-2021). Panels (b) and (d) show spatial differences in the MSE (averaged over time) of NN-GPR and LM for T2M(b) and PR(d).

Regional multi-model analysis

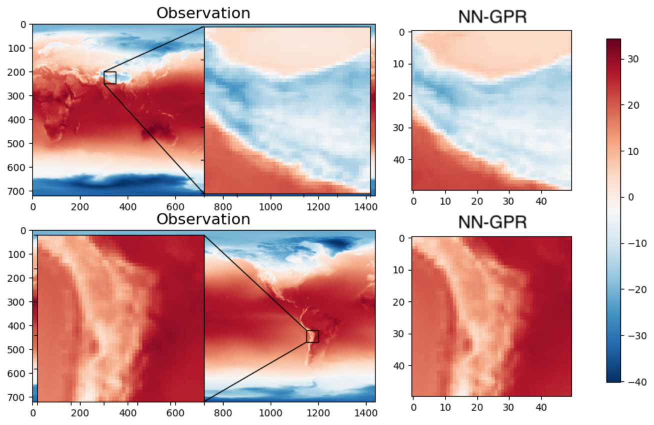


Figure 3: NN-GPR prediction vs the observed reanalysis field.

NN-GPR shows remarkable detail at the local level for a global model.
Automatic downscaling.

Part 2. Quantifying uncertainty

How to quantify uncertainty of f_θ? Without more assumptions?

- **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

Proposal: Conformal inference to construct exact prediction sets for f_θ

Given a confidence level $\alpha \in (0, 1)$ construct a non-trivial prediction sets $C_\alpha(X)$ such that

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

Conformal inference

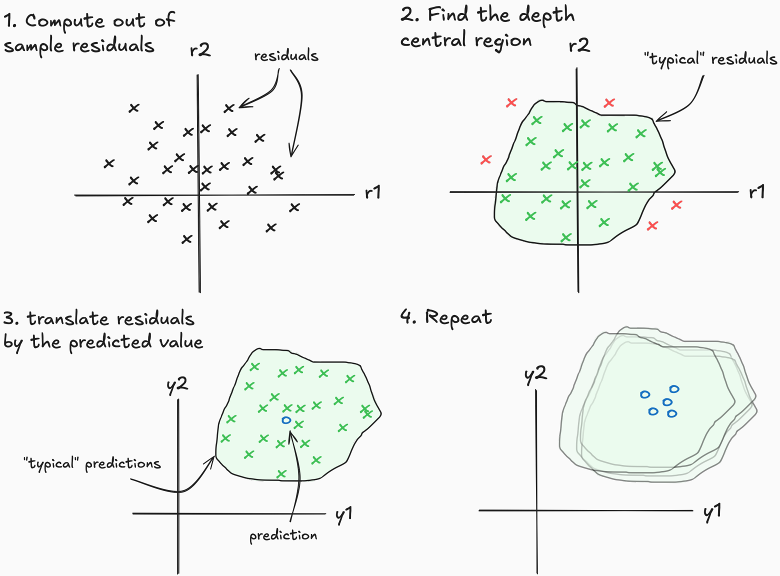
Conformal inference requires two functions to construct such a $C_\alpha(X_t)$

1. Ensemble analysis function $f_\theta : X_t \mapsto Y_t$
    - Converts the ensemble X_t into a "prediction" of Y_t
2. Scoring function $d : Y \times Y \mapsto \mathbb{R}$
    - ranks out-of-sample residuals, $Y_1 - \hat{Y}_1, ..., Y_N - \hat{Y}_N$ from "most typical" to "most outlying"
    - Identify a $(1-\alpha) \times 100\%$ typical set of residuals by retaining the $q = \lceil n+1 \rceil (1-\alpha)$ residuals with the lowest scores.

Data depth: quantifies centrality with respect to a distribution.

- Monotonically and (usually) continuously decrease from 1 (centroid) to 0 (infinitely outlying)
- Upper level sets $D_\beta(X) = \{z \in \mathbb{R}^d : D(z \mid X) \geq \beta\}$ contain all z with depths greater than β (more central than β)
- Upper level sets = typical sets (most central)

Algorithm – Conformal Central Region



1. Compute out of sample residuals

2. Find the depth central region

3. translate residuals by the predicted value

4. Repeat

Results - Coverage

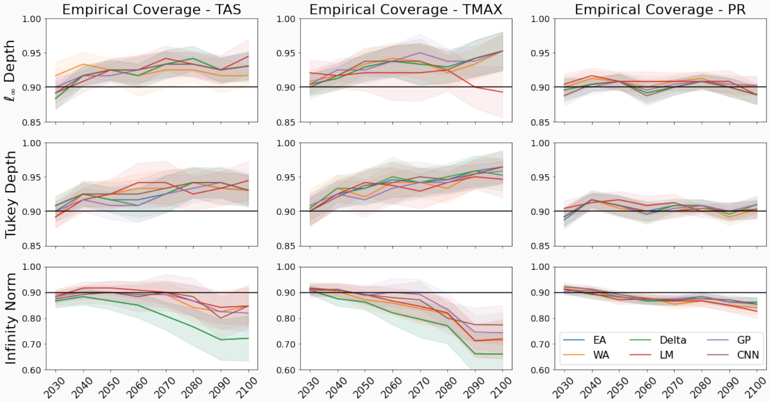


Figure 4: Empirical coverage of the ($\alpha = 0.1$) conformal prediction set generated by each depth function, for each ensemble analysis function and climatic variable. Proper depth functions (ℓ_∞-Depth and Tukey depth) show consistent nominal to slightly conservative coverage

Results - 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 3: 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.)

Results - Global projections

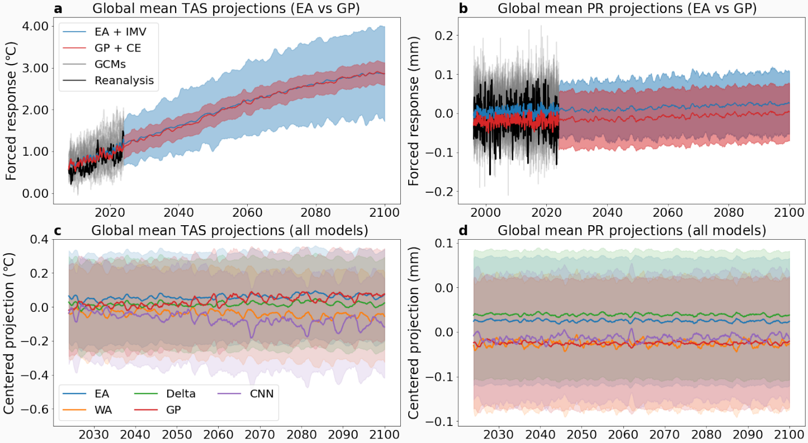


Figure 5: Panels (a) and (b) show the projected global mean monthly averages using EA + IMV and GP + CE. Panels (c) and (d) show the de-meaned projections, with CE based bands, for all models, excluding LM. Projections and uncertainty bands are smoothed using a 12 month moving average to reduce seasonal variability in the plot.

Summary

- Statistical learning / prediction is a simple and flexible approach for ensemble analysis
  - Part 1. Framework for implicitly distilling models into a single projection
  - Part 2. Framework for quantifying projection uncertainty via projection sets
- NNGPR model - Analysis function
  - Significant projection skill improvements over existing approaches, poor UQ.
  - Reproduces spatial details well enough to act as surrogate RCM
  - Semi-robust to covariate shift / distribution shift issues
- Conformal inference - UQ
  - Works for any ensemble analysis function and (proper) depth function
  - Exact to conservative coverage and Improved UQ skill compared to IMV and IMV(BC).
  - Generally improves marginal spatial UQ

Reference

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

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

Appendix: 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 (reverse) depths of the residual fields R_t with respect to $\mathcal{R}_{\text{cal}}$

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

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

Appendix: Metrics - calibration sensitivity

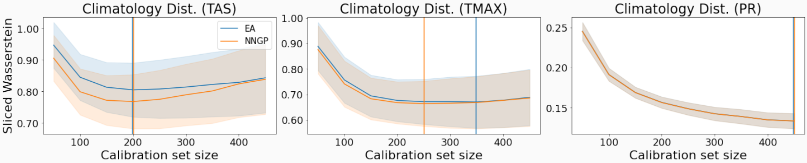


Figure 6: Blue lines show the mean, over all perfect model experiments, sliced Wasserstein distance between EA + CE and the target model over the test period (2024-2100) using an increasing large calibration set size. Orange line show the same except using GP + CE. Blue and orange shading represent ±2 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.

Appendix: Validation

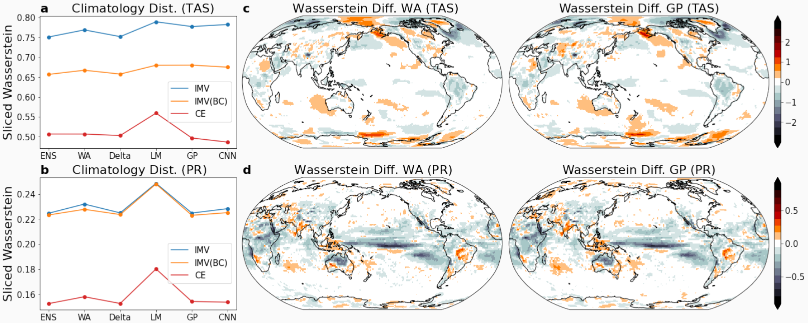


Figure 7: Panels (**a**) and (**b**) show the sliced Wasserstein distance computed between the projected distribution, for each model plus UQ combination, and the reanalysis data on the held out dataset for TAS and PR, respectively. Panels (**c**) and (**d**) show pointwise Wasserstein difference between 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.