MULTIMODEL ENSEMBLE PREDICTIONS OF PRECIPITATION USING BAYESIAN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodel ensembling improves predictions and considers model uncertainties. In this study, we present a Bayesian Neural Network (BNN) ensemble approach for large-scale precipitation predictions based on a set of climate models. BNN infers spatiotemporally varying model weights and biases through the calibration against observations. This ensemble scheme of BNN sufficiently leverages individual model skill for accurate predictions as well as provides interpretability about which models contribute more to the ensemble prediction at which locations and times to inform model development. Additionally, BNN accurately quantifies epistemic uncertainty to avoid overconfident projections. We demonstrate BNN's superior prediction performance to three state-of-the-art ensemble approaches and discuss its interpretability and uncertainty quantification.

1 INTRODUCTION AND MOTIVATION

Improving large-scale precipitation predictability is a complex problem and contingent on improving understanding of large-scale fluxes of water, energy, and biogeochemical cycles. Many climate models have been applied for precipitation simulations. They have large uncertainties in physical process representations, show varying prediction skills at different locations and times, and are usually not constrained by observations. Improving precipitation prediction requires a comprehensive multimodel ensemble approach that leverages each individual climate model's spatiotemporally varying predictive skill and integrates observations to reduce uncertainty using a calibration framework.

Existing ensemble approaches usually assume model independence and model democracy in which each model is weighted equally [1; 2; 3]. However, neither of these assumptions is true. Many climate models in Coupled Model Intercomparison Projects (CMIPs) share components or are variants of another model in the ensemble, and these models have large inconsistency in their skill at a given location and time. Even an individual climate model shows considerably inconsistent skills at different locations and times. By recognizing different capabilities among models, some studies assigned distinct weights to individual ensemble members [4; 5; 6; 7; 8; 9]. One of the most adopted ensemble weighting scheme was proposed by [10], which calculated model weights by balancing the model skill and model uniqueness. However, the determination of its trade-off parameters is subjective, and the values can greatly affect the model ensembling results [11; 12]. More importantly, existing weighted ensemble methods assign a uniform weight to a model across space and time and the same weight is applied for future projections. This may result in an inaccurate and overconfident precipitation prediction and thus misguide water management decision making.

In this work, we propose a Bayesian neural network (BNN) ensemble approach to improve large-scale precipitation predictability. This approach leverages machine learning power in data analytics and predictive analytics to calculate spatiotemporally varying model weights and biases by taking advantage of climate models' diverse performance in heterogeneous geography and different seasons. BNN uses observations to constrain the ensemble prediction while considering heteroscedastic data uncertainty. It will also consider extrapolation or epistemic uncertainty as we project to the future climate. Additionally, the proposed approach will provide interpretations and insights of the model and the data to inform the model and data development.

The main contributions of this work are:

1. We apply a BNN ensemble approach for large-scale precipitation prediction to improve predictive performance by leveraging individual model's spatiotemporally varying skill and calibrating the model weights and biases using observations.
2. We demonstrate the superior performance of BNN to three state-of-the-art methods and show that BNN can accurately calculate the epistemic uncertainty to avoid overconfident projection.
3. We provide interpretability about which models contribute more to the ensemble prediction at which locations and times and justify the superior performance of BNN.

2 Bayesian Neural Networks for Ensemble Model Prediction

We assume that observations $y(\mathbf{x}, t)$ at a given location $\mathbf{x}$ and time t can be modeled as a sum over an ensemble of m climate model predictions $M_i(\mathbf{x}, t)$ weighted by their respective weights $\alpha_i(\mathbf{x}, t)$, a bias term $\beta(\mathbf{x}, t)$, and a data noise term $\sigma(\mathbf{x}, t)$. BNN reads the data location ($\mathbf{x}$) and time (t) as inputs and estimates the model weights, biases, and data noises at the given $(\mathbf{x}, t)$ by calibrating its outputs against observations. As illustrated in Figure 1, BNN first uses a set of dense layers to extract common information of the model weights, biases, and data noises and then designs three sets of dense layers to learn the information specific to each component respectively. Next, BNN incorporates the multiple climate model predictions $M_i(\mathbf{x}, t)$ and combines them with the calculated model weights, biases, and data noises in the loss function to match the ensemble prediction $\hat{y}(\mathbf{x}, t) = \sum_{i=1}^{m} \alpha_i(\mathbf{x}, t) M_i(\mathbf{x}, t) + \beta(\mathbf{x}, t)$ with the observations.

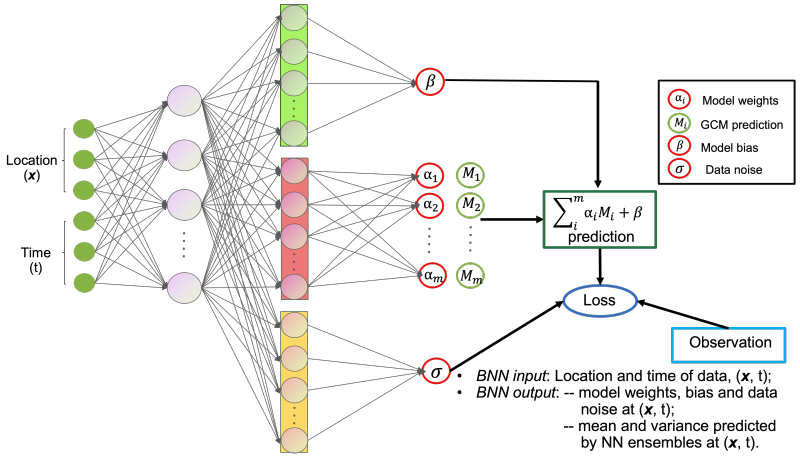

Figure 1: Architecture of Bayesian neural networks.

The network is trained in the Bayesian context using the randomized maximum a posteriori (MAP) sampling [13; 14]. The MAP sampling approach uses multiple neural networks to quantify the machine learning model prediction uncertainty. Specifically, for the j-th neural network, we draw a sample from the prior distribution over the network parameters (assumed Gaussian) $\theta_{anc,j} \sim N(\mu_{prior}, \Sigma_{prior})$, and compute the MAP estimate corresponding to a prior re-centered at $\theta_{anc,j}$. When we consider a dataset of N observations $\{y_i, \mathbf{x}_i, t_i\}$ and specify the data likelihood by assuming Gaussian noise $\sigma(\mathbf{x}, t)$, the calculation of the MAP estimate is equivalent to minimize the following loss function for the j-th neural network,

$$Loss_j = \sum_{i=1}^{N} \frac{(y_i - \hat{y}_j(\mathbf{x}_i, t_i))^2}{\sigma_j^2(\mathbf{x}_i, t_i)} + \sum_{i=1}^{N} log(\sigma_j^2(\mathbf{x}_i, t_i)) + ||\Sigma_{prior}^{-1/2}(\theta_j - \theta_{anc,j})||_2^2. \quad (1)$$

The prediction of a trained ensemble with n_e neural networks is thus a mixture of n_e Gaussians, $N(\hat{y}_j(\mathbf{x}_i, t_i), \sigma_j^2(\mathbf{x}_i, t_i))$. The mean value of these neural network predictions $\frac{1}{n_e}\sum_j \hat{y}_j$ will be the final BNN prediction results. The variance $\frac{1}{n_e}\sum_j \sigma_j^2 + \frac{1}{n_e}\sum_j \hat{y}_j^2 - (\frac{1}{n_e}\sum_j \hat{y}_j)^2$ quantifies predictive uncertainty where the first term quantifies data uncertainty and the second and third terms quantify epistemic uncertainty (see [13] for details).

3 PRECIPITATION DATA AND PREDICTION PROBLEM

We first design a simple precipitation prediction problem to validate BNN. The details of this example are described in Appendix A.1. The experiment results indicate that BNN can reasonably calculate spatially varying model weights by assigning the weights only in regions where they are skillful; and attributed to its advanced weighting scheme, BNN demonstrates accurate ensemble predictions.

We then apply BNN to Coupled Model Intercomparison Phase-6 (CMIP6) Global Climate Models (GCMs) for precipitation prediction. The precipitation data focus on United States at monthly scale in 1980-2014, downloaded from Earth System Grid Federation (ESGF) archives (https://esgf-node.llnl.gov/search/cmip6). For demonstrating BNN and evaluating its predictive performance, we consider seven GCMs in this study where six of them are for model predictions and the other one is selected as a synthetic truth for validation. The six individual models are Alfred Wegener Institute Climate Model (AWI-CM-1-1-MR), Manabe Climate Model v1.0 - University of Arizona (MCM-UA-1-0), Community Earth System Model Version 2 (CESM2), and three versions of Energy Exascale Earth System Model (E3SM-1-0, E3SM-1-1-ECA, E3SM-1-1-ECA). The synthetic truth is from the model CESM2-WACCM. All these GCMs precipitation data are remapped to a common 1° latitude-longitude grid. The distributions of these data are shown in Appendix A.2, where the three models AWI-CM-1-1-MR, MCM-UA-1-0, and CESM2 have similar precipitation with the synthetic truth. We use the first 20 years' data for training and the last 15 years for out-of-sample testing. The BNN simulation setup is described in Appendix A.3 in detail.

4 RESULTS AND DISCUSSION

In this section, we first evaluate BNN's prediction performance in comparison with three state-of-the-art ensemble approaches (introduced in Appendix A.4) and then assess BNN's interpretability by analyzing its spatiotemporally varying model weights and biases to interpret which models contribute more to the ensemble predictions at which locations and seasons.

Figure 2 shows the absolute prediction errors of the four ensemble approaches averaged over the simulation period. The figure indicates that BNN outperforms the other three approaches by producing smaller prediction errors in the entire domain. Although both the weighted average and spatially weighted average approaches use model-skill-based weights for ensemble prediction, it seems that their weighting schemes fail to bring significant improvement over the simple average method, where the root mean squared errors (RMSEs) are close to each other with values of 1.445, 1.443 and 1.452 for simple average, weighed average and spatially weighted average, respectively. Figure 3(a) summarizes spatial precipitation data averaged over time for the six individual models and the four ensemble predictions. It further demonstrates that BNN produces closer prediction results to the synthetic truth and there is no much difference between the other three ensemble approaches.

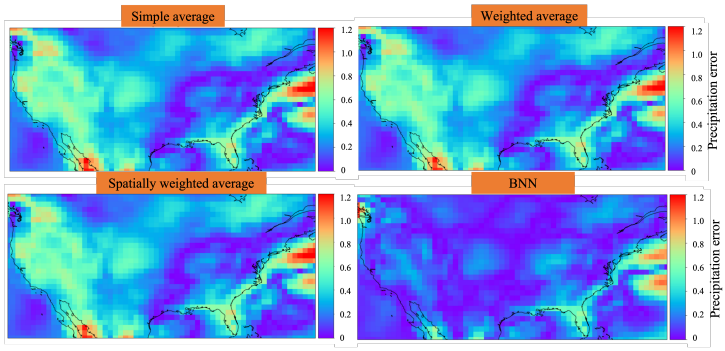

Figure 2: Absolute precipitation errors of the four ensemble approaches averaged over the simulation period. BNN outperforms the other three approaches by producing smaller prediction errors.

The superior prediction performance of BNN is partially attributed to its spatiotemporally varying weights. As shown in Figure 3(b) where we present the temporally averaged weights for the six models, the three top-performing models, AWI-CM-1-1-MR, MCM-UA-1-0, and CESM2, receive higher weights than the others. Additionally, the weights in each individual model vary spatially

where the regions having small prediction errors shows higher weights. This suggests that BNN sufficiently leverages models' geographically heterogeneous prediction skill.

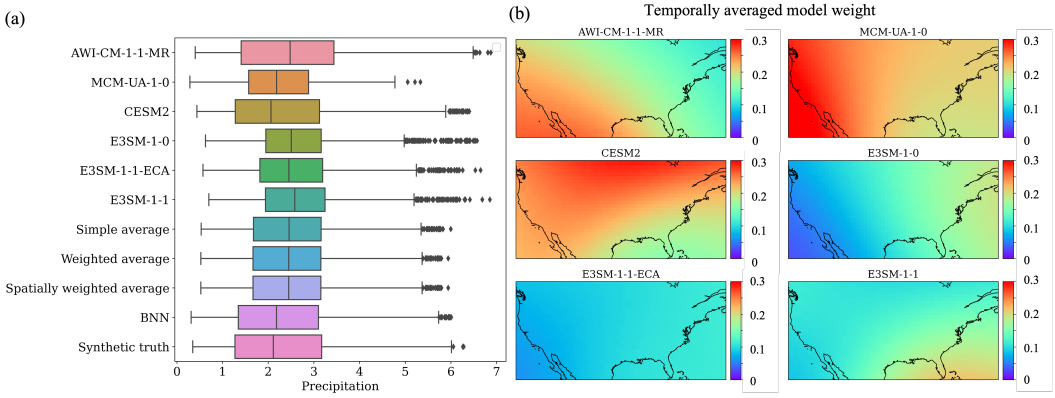

Figure 3: (a) Boxplot of the precipitation data averaged over simulation time for the six individual models and the four ensemble approaches; (b) Temporally averaged model weights of the six individual models.

Besides the smart weighting scheme, the spatially varying bias term in BNN also plays an important role for accurate precipitation prediction. As shown in Figure 4(a) which presents the weighted sum of the prediction errors of the six models, the northwest region has a relatively large positive prediction error. To compensate the error and make the ensemble prediction fit the calibration data well, BNN estimates the bias with a relatively large negative value in the region as depicted in Figure 4(b). This bias compensation scheme is particularly important when all the individual models generate overestimation or underestimation, in which case the ensemble prediction will hardly match well with the observations despite the ensembling schemes. Under this situation, by introducing the bias term and calibrating its value against the data, we can improve the ensemble predictions. In this example, we demonstrate that the BNN ensemble approach, by spatiotemporally identifying skillful models and calibrating the ensemble prediction against the data using the bias compensation, produces accurate predictions and shows better predictive performance than the state-of-the-art methods.

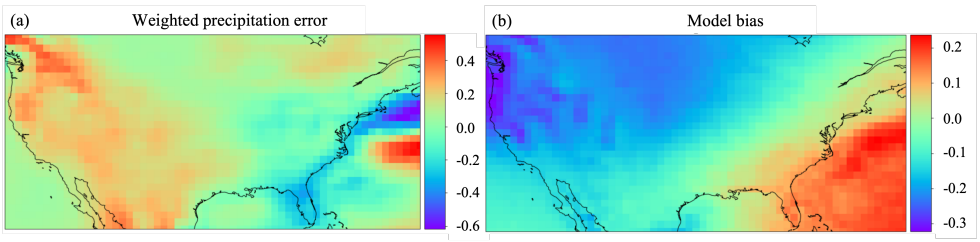

Figure 4: (a) Weighted precipitation errors of the six models; (b) Estimated bias (β in Figure 1) in BNN.

In addition to competence in accurate prediction, BNN also provides interpretability, allowing for identification of regions and seasons in which particular models contribute more to the ensemble prediction. We divide the simulation domain into four regions (see Appendix A.2). As illustrated in Figure 5(a), although models MCM-UA-1-0 and CESM2 have the highest weights globally, MCM-UA-1-0 contributes highly in the West and CESM2 is the dominant model in North and East regions. BNN accurately identifies the spatially skillful models. Take a close examination in the West region (Figure 5(b)), MCM-UA-1-0 receives higher weights than E3SM-1-0 because MCM-UA-1-0 performs better with smaller prediction errors. Figure 6(a) investigates BNN's temporally varying weights. All the models show seasonally changing weights and none of the models performs the best all the time. This suggests the importance of calculating temporally varying weights in the ensembling. Figure 6(b) demonstrates that BNN can accurately estimate the weights by leveraging the models' seasonally distinct skill. For example, in August 1991, model CESM2 shows smaller prediction errors than MCM-UA-1-0, and BNN assigns a higher weight to CESM2 at this time point. The uncertainty analysis of BNN is presented in Appendix A.5. Results suggest that BNN can accurately quantify the uncertainty to avoid overconfident future projection.

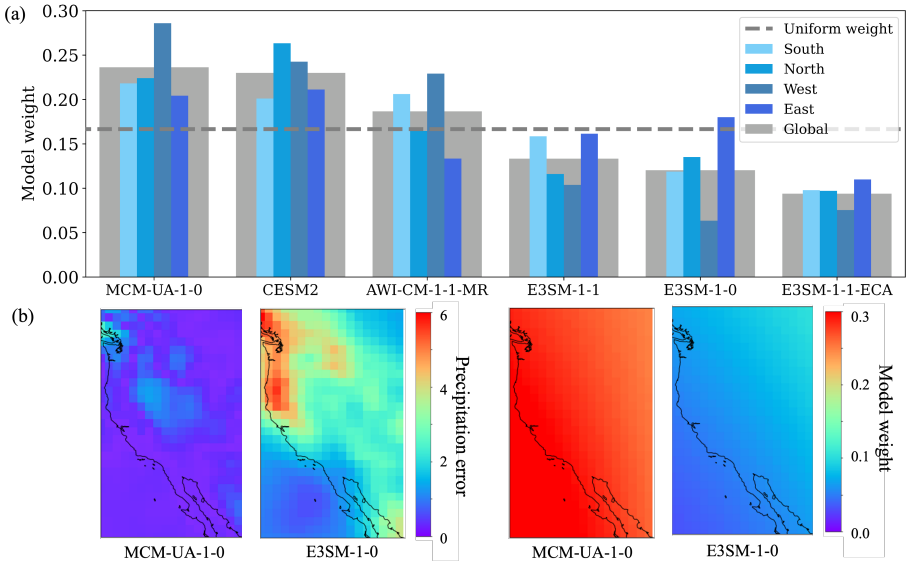

Figure 5: (a) Temporally averaged weights in the entire domain (Global) and the four regions (see Appendix A.2); (b) Prediction error and model weights of models MCM-UA-1-0 and E3SM-1-0 in the West region.

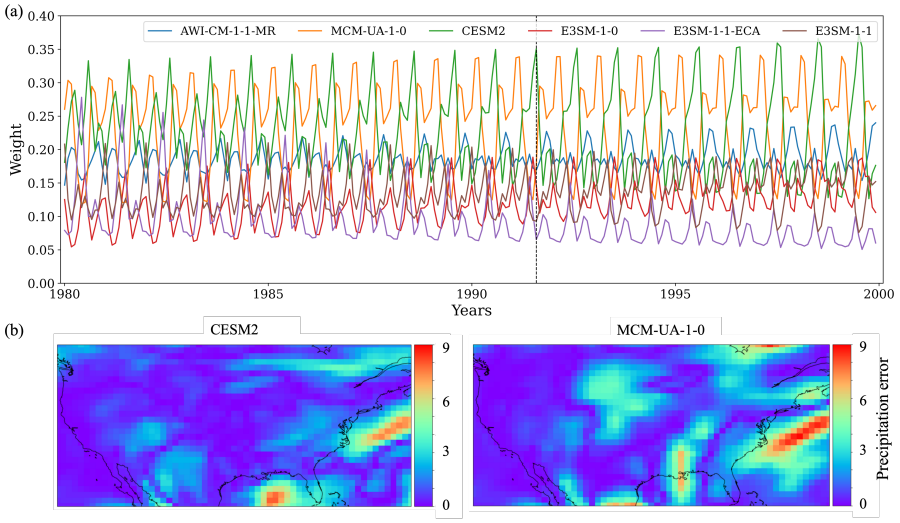

Figure 6: (a) Spatially averaged weights in the simulation period; (b) Prediction errors of models CESM2 and MCM-UA-1-0 in August 1991.

5 Conclusions and Future Work

We present a BNN ensemble approach for climate model predictions and demonstrate its superior performance to the three state-of-the-art methods in the application of predicting precipitation. First, BNN calculates spatiotemporally varying model weights to fully leverage individual model skill at different locations and seasons. Second, BNN estimates model biases by fitting to the observation data, which further improves the prediction performance. Third, BNN accurately quantifies the epistemic uncertainty which circumvents the overconfidence in future projection. Last but not the least, BNN has interpretability about which models contribute more to the ensemble prediction at which locations and times. This insight advances our understanding of each individual model's skill across the space and time and inform the model development. In the future, we will include more climate models in the ensemble analysis and apply BNN for different climate variable predictions.

REFERENCES

[1] P. J. Gleckler, K. E. Taylor, and C. Doutriaux, "Performance metrics for climate models," *Journal of Geophysical Research: Atmospheres*, vol. 113, no. D6, 2008.

[2] R. Knutti, R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, "Challenges in combining projections from multiple climate models," *Journal of Climate*, vol. 23, no. 10, pp. 2739–2758, 2010.

[3] R. Pincus, C. P. Batstone, R. J. P. Hofmann, K. E. Taylor, and P. J. Glecker, "Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models," *Journal of Geophysical Research: Atmospheres*, vol. 113, no. D14, 2008.

[4] M. Amos, P. J. Young, J. S. Hosking, J.-F. Lamarque, N. L. Abraham, H. Akiyoshi, A. T. Archibald, S. Bekki, M. Deushi, P. Jöckel, *et al.*, "Projecting ozone hole recovery using an ensemble of chemistry–climate models weighted by model performance and independence," *Atmospheric Chemistry and Physics*, vol. 20, no. 16, pp. 9961–9977, 2020.

[5] L. Brunner, R. Lorenz, M. Zumwald, and R. Knutti, "Quantifying uncertainty in european climate projections using combined performance-independence weighting," *Environmental Research Letters*, vol. 14, no. 12, p. 124010, 2019.

[6] R. Lorenz, N. Herger, J. Sedláček, V. Eyring, E. M. Fischer, and R. Knutti, "Prospects and caveats of weighting climate models for summer maximum temperature projections over north america," *Journal of Geophysical Research: Atmospheres*, vol. 123, no. 9, pp. 4509–4526, 2018.

[7] S. Wenzel, V. Eyring, E. P. Gerber, and A. Y. Karpechko, "Constraining future summer austral jet stream positions in the cmip5 ensemble by process-oriented multiple diagnostic regression," *Journal of Climate*, vol. 29, no. 2, pp. 673–687, 2016.

[8] A. Y. Karpechko, D. Maraun, and V. Eyring, "Improving antarctic total ozone projections by a process-oriented multiple diagnostic ensemble regression," *Journal of the Atmospheric Sciences*, vol. 70, no. 12, pp. 3959–3976, 2013.

[9] J. Räisänen, L. Ruokolainen, and J. Ylhäisi, "Weighting of model results for improving best estimates of climate change," *Climate dynamics*, vol. 35, no. 2, pp. 407–422, 2010.

[10] B. M. Sanderson, R. Knutti, and P. Caldwell, "A representative democracy to reduce interdependency in a multimodel ensemble," *Journal of Climate*, vol. 28, no. 13, pp. 5171–5194, 2015.

[11] R. Knutti, J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring, "A climate model projection weighting scheme accounting for performance and interdependence," *Geophysical Research Letters*, vol. 44, no. 4, pp. 1909–1918, 2017.

[12] B. M. Sanderson, M. Wehner, and R. Knutti, "Skill and independence weighting for multi-model assessments," *Geoscientific Model Development*, vol. 10, no. 6, pp. 2379–2395, 2017.

[13] U. Sengupta, M. Amos, J. S. Hosking, C. E. Rasmussen, M. Juniper, and P. J. Young, "Ensembling geophysical models with bayesian neural networks," *arXiv preprint arXiv:2010.03561*, 2020.

[14] T. Pearce, M. Zaki, A. Brintrup, N. Anastassacos, and A. Neely, "Uncertainty in neural networks: Bayesian ensembling," *stat*, vol. 1050, p. 12, 2018.

[15] J. Muñoz-Sabater, E. Dutra, A. Agustí-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, H. Hersbach, *et al.*, "Era5-land: A state-of-the-art global reanalysis dataset for land applications," *Earth System Science Data*, vol. 13, no. 9, pp. 4349–4383, 2021.

A APPENDIX

A.1 A SYNTHETIC CASE TO VALIDATE BNN APPROACH

We construct a synthetic case to validate the Bayesian neural network (BNN) ensemble scheme. The synthetic truth dataset is the ERA5 reanalysis precipitation data from the European Centre for Medium-Range Weather Forecasts reanalysis in period 1980-2014 [15]. The original ERA5 data is at 33 km horizontal grid spacing and hourly scale. We aggregate the data to monthly scale and remap it to a common 1° latitude-longitude grid. The rescaled data (Figure 7(a)) is used as the synthetic truth. We design four models to replicate the synthetic truth but only in distinct geographical regions, i.e., model I, II, III, and IV are correct only in region I, II, III, and IV, respectively (Figure 7(a)), and the other regions in each model are generated with random noise. We train BNN using the first 20 years of data and use the remaining 15 years for out-of-sample validation.

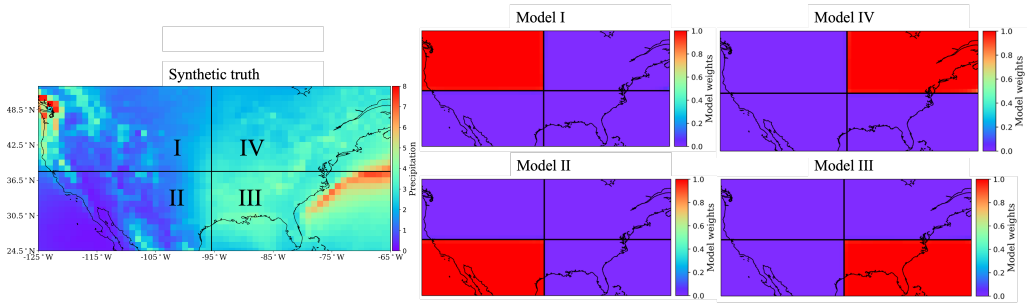

Figure 7: (a) Rescaled ERA5 precipitation data serving as a synthetic truth; (b) Model weights recovered by BNN. BNN only assigns weights to the regions where they are skillful.

Figure 7(b) indicates that BNN has successfully recovered the expected model weights; models only have weights in regions where they are skillful. Because BNN reasonably leverages each individual model's prediction skill by accurately calculating the spatially varying weights, the ensemble predictions of BNN have a great agreement with the synthetic truth (Figure 8). In contrast, the simple average method assigns uniform and equal weights to the models in the entire domain, which produces predictions dramatically different from the "truth". This example validates BNN's ability to successfully capture model skill and demonstrates its competence in accurate ensemble prediction.

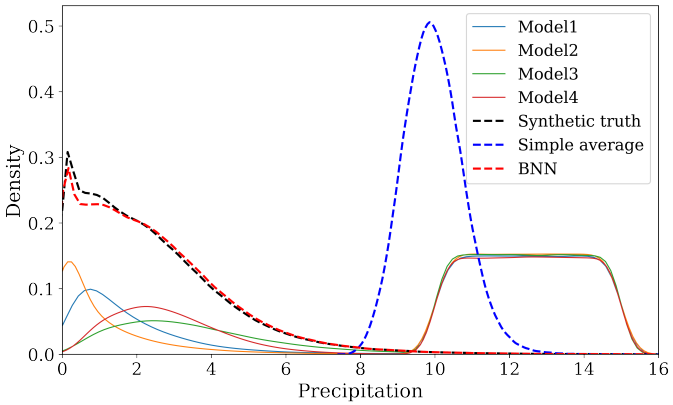

Figure 8: Probability density functions of the precipitation from the four individual models, the simple average and BNN ensemble approaches. The good agreement between BNN prediction and the synthetic truth demonstrates BNN's competence in accurate ensemble prediction.

A.2 PRECIPITATION DATA AND MODELS

The following seven CMIP6 GCMs are selected in this study for multi-model prediction analysis: Alfred Wegener Institute Climate Model (AWI-CM-1-1-MR), Manabe Climate Model v1.0 - University

of Arizona (MCM-UA-1-0), Community Earth System Model Version 2 (CESM2, CESM2-WACCM), and Energy Exascale Earth System Model (E3SM-1-0, E3SM-1-1-ECA, E3SM-1-1-ECA). The precipitation data focus on United States at monthly scale in 1980-2014. To demonstrate and validate BNN's capability, the data from model CESM2-WACCM is taken as synthetic truth. Its averaged prediction over the 35 years is shown in Figure 9(a), where we divide the domain into four regions for further analysis. The regional analysis results are illustrated in Figure 5 of the main text. The probability density functions (PDFs) of the precipitation data across the domain and the 35 years from the seven models are shown in Figure 9(b). The PDFs suggest that models AWI-CM-1-1-MR, MCM-UA-1-0, and CESM2 produce predictions closer to the synthetic truth than the three E3SM models. In Section 4 of the main text, we demonstrate that BNN can accurately assign higher weights to the three top-performing models.

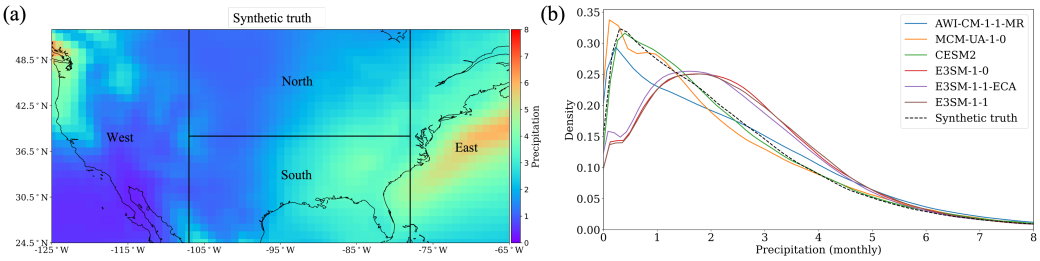

Figure 9: (a) Precipitation data of the synthetic truth averaged over 35 years; (b) probability density functions of the precipitation data from the six models for ensemble analysis and the synthetic truth.

A.3 Hyperparameters and prior values for the BNN simulation

To generate continuous model weights, biases, and noises that also respect seasonality, the location input $(\mathbf{x})$ is expressed by the Euclidean coordinate $(cos(lat)sin(lon), cos(lat)cos(lon), sin(lat))$ and the time input (t) is warped onto a 3D helix $(cos(2\pi t/T), sin(2\pi t/T), t)$, where $T = 1$ month. As illustrated in Figure 6 of the main text, the generated weights have an apparent annual periodicity and some variations over years. Additionally, both the spatial and temporal input variables are appropriately scaled to ensure the network outputs having desired geographic length scales and seasonal variations. Specifically, the spatial coordinates are scaled into the range [-2, 2], and the temporal coordinates (month of year and total months) are scaled into the range [-1, 1]. The BNN architecture is constructed with 50 neural network ensemble members and 100 nodes in each hidden layer. For the first hidden layer, the tanh activation function is used because of its zero-centered output. The softmax activation function is used for the model weight output layer to constrain the sum of model weights to a unity. The prior design is a key element to ensure BNN's accuracy in precipitation prediction. The output of the untrained softmax layer is initialized with a variance of 1.0 to ensure the feasibility of any model combinations at any locations and times. The prior variances for the bias and noise terms should be restricted to a small value to ensure the precipitation prediction mainly dependent on the combination of physical models. In this work, the bias mean and standard deviation are 0 and 0.01, and the noise mean and standard deviation are 0.02 and 0.004. We use the Adam optimizer with the batch size of 20000 and the learning rate of 0.0001. The training ends at the epoch of 1200.

A.4 State-of-the-art ensemble approaches

In this section, we introduce three state-of-the-art ensemble approaches which serves as baselines to evaluate BNN's performance.

A.4.1 Simple average

The simple ensemble averaging is to assign each model with equal weight. The prediction is given by

$$y(\mathbf{x}, t) = \frac{1}{N} \sum_{i=1}^{N} M_i(\mathbf{x}, t) \tag{2}$$

A.4.2 WEIGHTED AVERAGE

Based on the concepts proposed by Sanderson et al. [10], Knutti et al. [11] presented an ensemble weighting scheme to project the future sea ice change in the Arctic in consideration of both model skill and model independence. For an ensemble of N models, the weight w_i for model i is given by

$$w_i = \exp\big(-\frac{D_i}{\sigma_D^2}\big)/\big(1+\sum_{j\neq i}^{N}\exp\big(-\frac{S_{ij}^2}{\sigma_S^2}\big)\big) \tag{3}$$

where D_i^2 is the distance between a model and the observation and S_{ij}^2 is the distance between two models. The fundamental characteristics of this method are that models in better agreement with observations get higher weights and models with high similarities get less weights. Therefore, two constants (σ_D and σ_S) representing model skill and model uniqueness in this weighing scheme have a significant influence on the ensemble weights. When σ_D gets a small value, only a small number of models are assigned with weights, whereas when σ_D gets a large value, the ensemble averaging converges to simple averaging with equal weights.

A.4.3 SPATIALLY WEIGHTED AVERAGE

For an ensemble of N models, the spatially weighed average is defined by

$$w_i = \exp\big(-\frac{D_i}{n_i\sigma_D^2}\big)/\big(1+\sum_{j\neq i}^{N}\exp\big(-\frac{S_{ij}^2}{n_i\sigma_S^2}\big)\big) \tag{4}$$

This equation differs from Equation 3 through the addition of n_i, which is the number of observations for all locations instead of averaging data over space.

A.5 UNCERTAINTY ANALYSIS OF BNN ENSEMBLE PREDICTION

The BNN ensemble approach can quantify both the data uncertainty and the epistemic uncertainy due to the lack of data. Figure 10 shows cumulative density function (CDF) of the epistemic uncertainty for the training and unseen test data. The figure indicates that BNN can reasonably quantify the uncertainty, where the epistemic uncertainty of the out-of-sample test data is greater than that of the training data. This is highly desirable behavior and crucial in practice to prevent overconfident extrapolation.

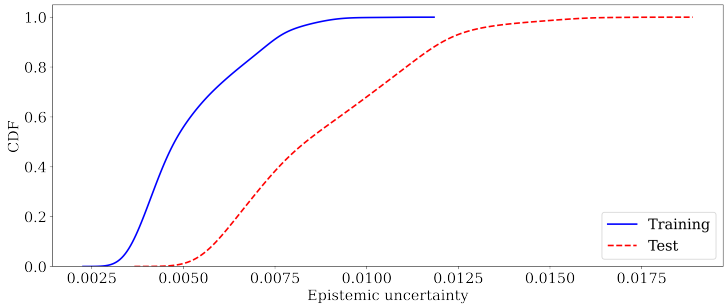

Figure 10: Epistemic uncertainty of the training and out-of-sample test data.