ENVIRONMENTAL RESEARCH
LETTERS

PURPOSE-LED PUBLISHING™

Part of the ENVIRONMENTAL RESEARCH SERIES

LETTER • OPEN ACCESS

Refined climatologies of future precipitation over High Mountain Asia using probabilistic ensemble learning

To cite this article: Kenza Tazi *et al* 2025 *Environ. Res. Lett.* **20** 084045

View the article online for updates and enhancements.

You may also like

- Precipitation phase drives seasonal and decadal snowline changes in high mountain Asia
M Bernat, E S Miles, M Kneib et al.

- Climate change impacts and adaptation to permafrost change in High Mountain Asia: a comprehensive review
Prashant Baral, Simon Allen, Jakob F Steiner et al.

- General overestimation of ERA5 precipitation in flow simulations for High Mountain Asia basins
He Sun, Fengge Su, Tandong Yao et al.

This content was downloaded from IP address 194.66.0.110 on 23/07/2025 at 15:53

IOP Publishing Environ. Res. Lett. 20 (2025) 084045 https://doi.org/10.1088/1748-9326/ade4dd

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
21 February 2025

REVISED
13 May 2025

ACCEPTED FOR PUBLICATION
16 June 2025

PUBLISHED
21 July 2025

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

LETTER

Refined climatologies of future precipitation over High Mountain Asia using probabilistic ensemble learning

Kenza Tazi[1,2,*], Sun Woo P Kim[3], Marc Girona-Mata[1,2] and Richard E Turner[1]

[1] Department of Engineering, University of Cambridge, Cambridge, United Kingdom
[2] British Antarctic Survey, Cambridge, United Kingdom
[3] Department of Physics, King's College London, London, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: kt484@cam.ac.uk

Keywords: High Mountain Asia, precipitation, ensemble learning, regional climate models, probabilistic, Gaussian process, CORDEX-WAS

Abstract

High Mountain Asia (HMA) holds the highest concentration of frozen water outside the polar regions, serving as a crucial water source for more than 1.9 billion people. Precipitation represents the largest source of uncertainty for future hydrological modelling in this area. In this study, we propose a probabilistic machine learning framework to combine monthly precipitation from 13 regional climate models developed under the Coordinated Regional Downscaling Experiment (CORDEX) over HMA via a mixture of experts (MoE). This approach accounts for seasonal and spatial biases within the models, enabling the prediction of more faithful precipitation distributions. The MoE is trained and validated against gridded historical precipitation data, yielding 32% improvement over an equally-weighted average and 254% improvement over choosing any single ensemble member. This approach is then used to generate precipitation projections for the near future (2036–2065) and far future (2066–2095) under RCP4.5 and RCP8.5 scenarios. Compared to previous estimates, the MoE projects wetter summers but drier winters over the western Himalayas and Karakoram and wetter winters over the Tibetan Plateau, Hengduan Shan, and South East Tibet.

1. Introduction

High Mountain Asia (HMA) stores more frozen water than anywhere else in the world after the Arctic and Antarctic polar caps. Its glaciers and snowfields, shown in figure 1, provide fresh water to more than 1.9 billion people through Asia's main rivers (Immerzeel *et al* 2020). Despite the critical importance of these water resources, much remains unknown about their distribution and how they will evolve under climate change. Among all hydrological drivers, precipitation contributes the greatest uncertainty to modelling future water security over HMA (Nie *et al* 2021, Orr *et al* 2022, Wester *et al* 2023). The main tools for understanding future precipitation over HMA are regional climate models (RCMs) (Maussion *et al* 2014, Palazzi *et al* 2015, Norris *et al* 2017, 2019, 2020, Orr *et al* 2017). However, complex orography and the lack of *in situ* hydrological observations, combined with limited RCM resolution and parametrisations over mountains, make precipitation difficult to model in this region (Orr *et al* 2017, Girona-Mata *et al* 2024, Tazi *et al* 2024, 2025). To overcome these challenges and understand the range of possible changes in precipitation, RCMs with different model physics and parametrisations are run, and their outputs combined.

Such collections of climate models are known as multi-model ensembles. In general, multi-model ensembles improve the accuracy of historical predictions and provide a way to explore the uncertainties in model-based climate projections (Kharin and Zwiers 2002, Palmer *et al* 2005, Stainforth *et al* 2007). However, the way in which models are combined is controversial (Tebaldi and Knutti 2007). This includes disagreements about how to measure aggregate accuracy (Eyring *et al* 2019), deal with ensemble outliers (Sanderson and Knutti 2012),

© 2025 The Author(s). Published by IOP Publishing Ltd

account for the proximity of model design (Knutti 2010, Masson and Knutti 2011), and address an infinite number of possible ensemble members (Tél et al 2020). As a result, the Intergovernmental Panel on Climate Change opted for the unweighted average of the global climate models (GCMs) from the Climate Model Intercomparison Project (CMIP) in their previous reports, using model spread as the uncertainty for future predictions (Flato et al 2014)[4].

In this paper, we present a probabilistic machine learning framework with the aim of combining precipitation outputs from 13 RCMs developed under the Regional Downscaling Experiment (CORDEX) for CMIP5. This method leverages the RCMs' heterogeneous seasonal and spatial biases to generate precipitation probability distributions for any location and month. This is achieved through a two-step approach. First, we compute statistical surrogates of the RCMs, i.e. a representation of their simulated precipitation field. For a given climatological period, we compile the empirical precipitation distribution for each RCM, location, and month. We then model the empirical distributions with Gaussian processes (GPs). GPs have been used in previous research for similar problems, such as spatiotemporal aggregation of GCMs using deep kernel learning (Harris et al 2013) or to find the latent signal of a temperature change in GCM ensembles for a single location (Amos et al 2022).

Second, we aggregate the RCM surrogate predictions at each spatiotemporal point via a mixture of experts (MoE), where distributions of experts (i.e. RCM surrogates) are combined using weights. We parametrise the weight of each expert by their Wasserstein distance with respect to historical data (Asian Precipitation-Highly Resolved Observational Data Integration Towards Evaluation of Water Resources or APHRODITE) and a learnt 'statistical temperature'. Our ensemble learning method is therefore similar to Bayesian model averaging (Min et al 2007, Olson et al 2016, Massoud et al 2023), where the similarity of the probability distribution with respect to historical data is used as a measure of faithfulness. However, unlike previous ensembling techniques, our method neither gives each RCM one vote, nor does it completely discard outliers with less accurate distributions; it only down-weighs them. As more experts are added, more precipitation events are considered, and the final distributions are refined.

Our main goal is to compare this method with that of Sanjay et al (2017). Their work presented CORDEX CMIP5 runs over HMA. We study the same domain, time periods, and RCMs using both a learnt MoE and an equally-weighted (EW) RCM aggregation scheme. The latter is equivalent to the ensembling method used in Sanjay et al (2017). The paper is structured as follows. The study area and data are first described in section 2. The method is then summarised in section 3 and validated in section 4. Predictions for a historical reference period (1976–2005), the near future (2036–2065), and the far future (2066–2095) for RCP4.5 and RCP8.5 are analysed in section 5. Finally, limitations and further work are discussed in section 6.

[4] The sixth report (AR6) discards 'hot-models' from CMIP6 that have been shown to be overly sensitive to emission forcing (Zelinka et al 2020, Lee et al 2021, Hausfather et al 2022).

2. Study area and datasets

2.1. Study area

Following Sanjay et al (2017), we study precipitation over HMA between 20°–40° N 60°–100° E as shown in figure 1 and three mountain subregions: the West Himalaya and Karakoram (HMA1, 32°–39° N 71°–79° E), the Central and East Himalaya (HMA2, 27°–32° N 76°–93° E), and the Hengduan Shan and Southeast Tibet (HMA3, 27°–32° N 93°–103° E). In addition to including most glaciers for HMA, these subregions are also representative of key synoptic atmospheric patterns that drive precipitation in this area. Precipitation over HMA1 is strongly influenced by the western disturbances during winter (December to March), while precipitation distributions over HMA2 and HMA3 are mainly driven by the Indian summer monsoon and the East Asian summer monsoon, respectively, from June to September. The study area also overlaps the Hindu Kush Himalayan (HKH) region. As defined by the International Centre for Integrated Mountain Development, the HKH extends from 16° to 39° N and 61° to 105° E and encompasses more than 4192 000 km^2 of mountains over Afghanistan, Bangladesh, Bhutan, China, India, Myanmar, Nepal, and Pakistan (Bajracharya and Shrestha 2011).

2.2. CORDEX-WAS

CORDEX is a global initiative that aims to provide high-resolution climate data for regional and local applications. CORDEX for the West Asia domain (CORDEX-WAS, −20°–50° N 20°–115° E), also known as the South Asia domain, uses RCMs to dynamically downscale global climate simulations from CMIP5 over most of HMA (Taylor et al 2012, Sanjay et al 2017). CORDEX-WAS outputs are chosen over other RCM simulations for their accessibility and rigorous experimental protocol which allow for the straightforward comparison of many models (Giorgi and Gutowski Jr 2015). CORDEX-WAS is made up of 13 models using three RCMs and ten driving CMIP5 GCMs, as listed in table 1.

The CORDEX outputs are stored on a 0.44° rotated grid (approximately 50 km resolution). However, for simplicity, we use interpolated outputs on a 0.5° regular grid. For precipitation, the Centre for Climate Change Research at Indian Institute of Tropical Meteorology used bilinear interpolation to

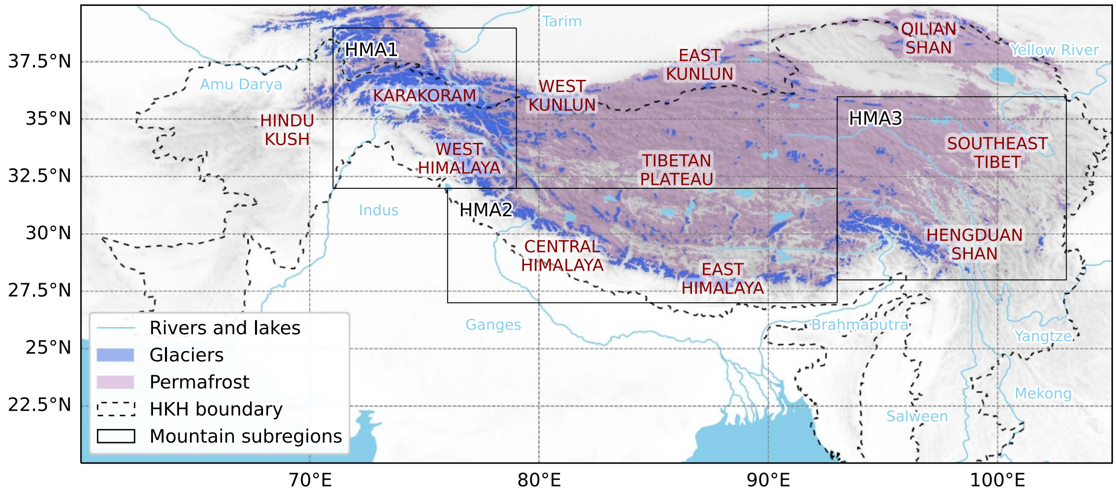

Figure 1. Map of High Mountain Asia showing major rivers and lakes (Lehner and Grill 2013), glaciers (RGI Consortium 2017), permafrost (Westermann *et al* 2024), the HKH boundary as defined by ICIMOD (2008), and three mountain subregions used for this study: the West Himalaya and Karakoram (HMA1), the Central and East Himalaya (HMA2), the Hengduan Shan and Southeast Tibet (HMA3). The map also includes the standardised names of the glacierised mountains regions (GTN-G 2023) referenced in this paper.

Table 1. CORDEX-WAS RCMs with CMIP5 driving models. Note that GFDL-ESM2M, MPI-ESM-LR, and CNRM-CM5 each drive two RCMs.

RCM	Driving CMIP5 model
IITM-RegCM4 (Giorgi *et al* 2012)	CanESM2 (Chylek *et al* 2011) GFDL-ESM2M (Dunne *et al* 2012) CNRM-CM5 (Voldoire *et al* 2013) MPI-ESM-MR (Jungclaus *et al* 2013) IPSL-CM5A-LR (Dufresne *et al* 2013)
SMHI-RCA4 (Samuelsson *et al* 2015)	CSIRO-Mk3.6 (Jeffrey *et al* 2013) EC-Earth (Hazeleger *et al* 2010) MIROC5 (Watanabe *et al* 2010) MPI-ESM-LR (Jungclaus *et al* 2013) IPSL-CM5A-MR (Dufresne *et al* 2013) GFDL-ESM2M CNRM-CM5[a]
MPI-REMO2009 (Teichmann *et al* 2013)	MPI-ESM-LR[b]

[a] The SMHI-RCA4 CNRM-CM5 model output only extends to 2085 for RCP8.5.
[b] MPI-REMO2009 MPI-ESM-LR model output only begins from 1961.

regrid the RCM data (Sanjay Jayanarayanan, Scientist 2025). We analyse CORDEX-WAS monthly precipitation outputs over 20°–40° N and 60°–105° E for historical (1950–2005) and projected (2006–2100) scenarios. Historical scenario experiments use estimates of greenhouse gas concentrations, aerosols, and land use (from observed or reconstructed data) to drive the GCMs. Projected scenario experiments are based on representative carbon pathways (RCPs), i.e. theoretical future emissions scenarios. In this paper, outputs for RCP4.5 and RCP8.5 are investigated. RCP4.5 represents a stabilisation pathway, where emissions peak around 2040 and then decline, while RCP8.5 represents a high-emissions pathway with minimal policy intervention. Going forward, we simplify simulation names to 'GCM RCM', e.g. the IITM-RegCM4 simulation driven by IPSL-CM5A-LR becomes IPSL RegCM4.

2.3. APHRODITE

APHRODITE is a gridded precipitation dataset ranging from 1951 to 2015 with a spatial resolution of 0.25° (Yatagai *et al* 2012). APHRODITE has one of the best spatiotemporal coverages over HMA and is one of the most studied and accurate precipitation products for the region (Dimri 2021). The dataset was created through the interpolation of precipitation gauge observations using a custom correlation distance lookup table. We use the monthly precipitation product APHRO_V1101 specifically developed

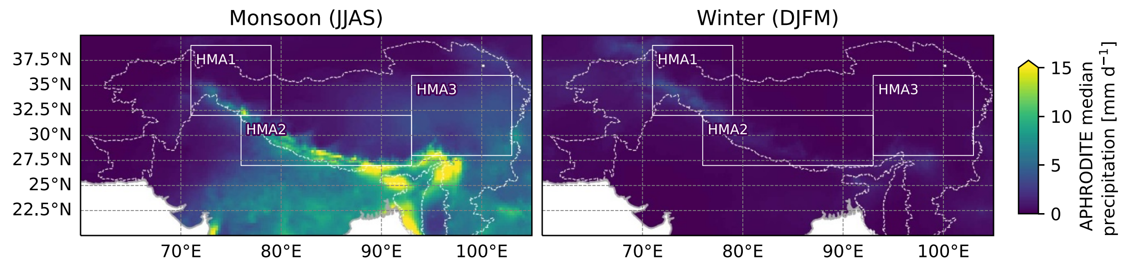

Figure 2. APHRODITE precipitation medians between 1976 and 2005 for the summer monsoon (left) and winter (right).

for monsoon Asia. Figure 2 plots APHRODITE precipitation between 1976 and 2005.

3. Method

This section presents our ensemble learning method, summarised in figure 3.

3.1. RCM surrogate models

3.1.1. GPs

To create the surrogate models for each RCM, we consider a regression problem with a dataset of N paired inputs $\boldsymbol{X} := \{\boldsymbol{x}_i\}_{i=1}^N$ and noisy observations $\boldsymbol{Y} := \{y_i\}_{i=1}^N$, where i denotes the ith data pair. Each input $\boldsymbol{x}_i$ is a three-dimensional vector of $(\text{month}, \text{latitude}, \text{longitude})$. Each observation y_i is a scalar, corresponding to a prediction of precipitation by an RCM.

The problem is modelled as $y_i = f(\boldsymbol{x}_i) + \epsilon_i$, where we assume that the noise $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma_\mathrm{n}^2)$ is independently and identically distributed. The function f can be represented with a GP (Rasmussen and Williams 2006). GPs are powerful non-parametric Bayesian models. A GP is a distribution over functions for which every finite set of outputs is jointly Gaussian distributed. GPs are fully specified by a mean function $\mu(\cdot)$ and a kernel function $k(\cdot,\cdot)$, which depend on hyperparameters θ_μ and θ_k, jointly referred to as θ. If we assume f to be a GP, i.e. $f \sim \mathcal{GP}(\mu(\cdot), k(\cdot,\cdot))$, it follows that $f(\boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{X}}, \mathbf{K}_{\boldsymbol{X}} + \sigma_\mathrm{n}^2\mathbf{I})$ where $(\boldsymbol{\mu}_{\boldsymbol{X}})_i := \mu(\boldsymbol{x}_i)$ is a N-dimensional vector and $(\mathbf{K}_{\boldsymbol{X}})_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is an $N \times N$ covariance matrix.

A posterior GP conditioned on the dataset $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{Y})$ can be inferred. This joint probability distribution of precipitation is updated using all data at every month and location of the RCM simultaneously. The GP is then trained by learning the optimal hyperparameters from the data. This optimisation is performed by maximising the log-marginal likelihood of the model

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{\operatorname{argmax}}\left[\log p\left(\boldsymbol{Y}|\boldsymbol{X},\theta\right)\right] \\ &= \underset{\theta}{\operatorname{argmax}}\left[\log \mathcal{N}\left(\boldsymbol{Y}|\boldsymbol{\mu}_{\boldsymbol{X}}(\theta_\mu), \mathbf{K}_{\boldsymbol{X}}(\theta_K) + \sigma_\mathrm{n}^2\mathbf{I}\right)\right].\end{aligned} \quad (1)$$

Finally, the GP can be sampled repeatedly at multiple arbitrary inputs. For a given input at an arbitrary month, latitude, and longitude, i.e. a test point $\boldsymbol{x}_*$, not necessarily seen in the training data, the mean and variance of the posterior predictive distribution of $f(\boldsymbol{x}_*)$ are given by

$$\underset{f|\mathcal{D}}{\mathbb{E}}\left[f(\boldsymbol{x}_*)\right] = \mu(\boldsymbol{x}_*) + \boldsymbol{K}_*^T\left(\mathbf{K}_{\boldsymbol{X}} + \sigma_\mathrm{n}^2\mathbf{I}\right)^{-1}\left(\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{X}}\right), \quad (2)$$

$$\underset{f|\mathcal{D}}{\operatorname{Var}}\left[f(\boldsymbol{x}_*)\right] = k(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{K}_*^T\left(\mathbf{K}_{\boldsymbol{X}} + \sigma_\mathrm{n}^2\mathbf{I}\right)^{-1}\boldsymbol{K}_*, \quad (3)$$

where $\boldsymbol{K}_* = k(\boldsymbol{X}, \boldsymbol{x}_*)$ is an N-dimensional vector.

The GP is a joint probability distribution of precipitation updated using all data at every month and location of the RCM simultaneously. Once trained, the GP can be sampled repeatedly at multiple arbitrary inputs.

3.1.2. Warped GPs

In its original formulation, the posterior distribution of the GP is normal. However, monthly precipitation y is generally log-normally distributed (Tazi *et al* 2024). Warped GPs handle non-normality in the data by transforming the target variables and GP outputs via a warping function (Snelson *et al* 2003). For this problem, we use a Box-Cox function $g_\lambda(\cdot)$ as the warping function. The warped outputs $\tilde{y}$ are then given by

$$\tilde{y} = g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{for} \quad \lambda \neq 0, \\ \log y & \text{for} \quad \lambda = 0, \end{cases} \quad (4)$$

where precipitation $y > 0$ is assumed to be positive and λ is the scaling factor. The scaling factor is optimised to make the distribution of $g_\lambda(y)$ as Gaussian as possible by maximising the log-likelihood of the model parameters arising from placing a Gaussian distribution over the transformed observations (Box and Cox 1964). In this study, we use a different optimised scaling factor λ_r for each RCM r

Finally, to recover the skewness and more complex distributional characteristics of the RCM data, the corresponding inverse transform is applied to the GP outputs. Henceforth, we will use $f(\cdot)$ to denote the

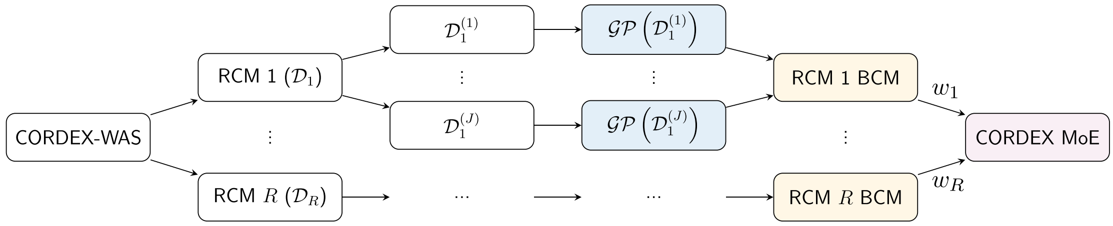

Figure 3. Ensemble learning method. For a given emission scenario and climatological period, each of the R RCMs outputs from CORDEX-WAS are split into J manageable spatiotemporal domains, $\mathcal{D}_r^{(j)}$ for the jth domain and rth RCM. A GP is fit to each domain and then combined using a BCM. The outputs of each RCM BCM are then combined using a weighted mixture model or mixture-of-experts (MoE) with weights $\{w_r\}_{r=1}^{R}$.

function that models the *warped* output $\tilde{y}$. We will denote the unwarped function, which may have heavy tails in its distribution, as $f_\lambda(\cdot) := g_\lambda^{-1}(f(\cdot))$. Details on the distribution of the unwarped precipitation can be found in appendix B.

3.1.3. Chained GPs

For this application, the variance of precipitation y should be considered as intrinsic climatological variance rather than noise. The GP output should therefore be heterogeneous in its variance, i.e. heteroskedastic. To address these requirements, we use chained GPs (Saul *et al* 2016) to model the variance and mean as latent functions such that

$$p(\tilde{y}_i | f_1, f_2, \boldsymbol{x}_i) = \mathcal{N}\left(\tilde{y}_i \,\middle|\, f_1(\boldsymbol{x}_i), \alpha(f_2(\boldsymbol{x}_i))\right), \tag{5}$$

where $f_1 \sim \mathcal{GP}(0, k_1(\cdot,\cdot))$, $f_2 \sim \mathcal{GP}(0, k_1(\cdot,\cdot))$ and α is a function used to map the outputs of f_2 to positive values. This mapping is required as the output of α represents the variance of a Gaussian distribution. Here, we apply an exponential function $\alpha(f_2(\boldsymbol{x}_i)) = \mathrm{e}^{f_2(\boldsymbol{x}_i)}$, but other functions, such as softplus, could be used. For our work, we use a shortcut to calculate the variance at each point without using further approximations, such as inducing points (Titsias 2009), by exploiting the gridded nature of our data.

3.1.4. Bayesian committee machines (BCMs)

Due to GP's poor scalability to large datasets, we use robust BCMs as a way of combining multiple GPs and speeding up inference (Tresp 2000, Deisenroth and Ng 2015, Cohen *et al* 2020). To do this, the training dataset is partitioned into J subsets of size M (so $N = JM$) where $M \ll N$. We define these subsets or domains as $\mathcal{D}^{(j)} = (\boldsymbol{X}^{(j)}, \boldsymbol{Y}^{(j)})$ for $j = 1, \ldots, J$. GPs are then used to model each domain. The robust BCM GPs share kernel hyperparameters, stopping individual experts from overfitting their local subset of data. Assuming conditional independence between domains $\mathcal{D}^{(j)}$ and by repeated application of Bayes' theorem, the following predictive distribution is obtained by combining the separate GP models:

$$p(f(\boldsymbol{x}_*) | \mathcal{D}) = \frac{\prod_{j=1}^{J} p_j^{\beta_j(\boldsymbol{x}_*)}\left(f(\boldsymbol{x}_*) | \mathcal{D}^{(j)}\right)}{p^{-1+\sum_j \beta_j(\boldsymbol{x}_*)}\left(f(\boldsymbol{x}_*)\right)}, \tag{6}$$

where $p(f(\boldsymbol{x}_*))$ is the prior distribution of the GP evaluated at test point $\boldsymbol{x}_*$ and β_j controls the contribution of expert j (see appendix A). The robust BCM's predictive mean and precision (inverse of the variance) are given by

$$\mu_*^{\mathrm{rbcm}} := \mathop{\mathbb{E}}_{f|\mathcal{D}^{(1)}\ldots\mathcal{D}^{(j)}}[f(\boldsymbol{x}_*)] = \left(\sigma_*^{\mathrm{rbcm}}\right)^2 \sum_{j=1}^{J} \beta_j(\boldsymbol{x}_*)\left(\sigma_*^{j}\right)^{-2} \mu_*^{j}, \tag{7}$$

$$\left(\sigma_*^{\mathrm{rbcm}}\right)^{-2} := \mathop{\mathrm{Var}}_{f|\mathcal{D}^{(1)}\ldots\mathcal{D}^{(j)}}[f(\boldsymbol{x}_*)]^{-1} = (1-M)\left(\sigma_*^{\mathrm{prior}}\right)^{-2} + \sum_{j=1}^{J} \beta_j(\boldsymbol{x}_*)\left(\sigma_*^{j}\right)^{-2}, \tag{8}$$

respectively, where μ_*^j and $(\sigma_*^j)^{-2}$ are the mean and precision of the predictive posterior given domain $\mathcal{D}^{(j)}$, and $(\sigma_*^{\mathrm{prior}})^{-2}$ is the prior precision of $p(f_*)$. This setup conserves the properties of an exact GP while improving scalability, moving from $\mathcal{O}(N^3)$ for training and $\mathcal{O}(N^2)$ for prediction to $\mathcal{O}(JM^3)$ and $\mathcal{O}(JM^2)$ respectively (Deisenroth and Ng 2015). For our work, the domains are divided temporally, such that each domain contains precipitation for a particular month over the whole spatial area.

With these adaptations, GPs can be used to interpolate large empirical distributions. To compare the RCM surrogates with the APHRODITE dataset introduced in section 2.3, we evaluate the surrogate GPs at the APHRODITE grid points. In figure 4, we illustrate the typical surrogate output with CSIRO RegCM4 for February. The surrogate distributions closely match the empirical RCM histograms, which can substantially differ with time and location.

3.2. Surrogate aggregation

3.2.1. MoE

The RCM surrogates are aggregated via the weighted sum of their posterior probability distributions. We refer to this weighted mixture model as a MoE. The resulting distribution for a given test point $\boldsymbol{x}_*$ is

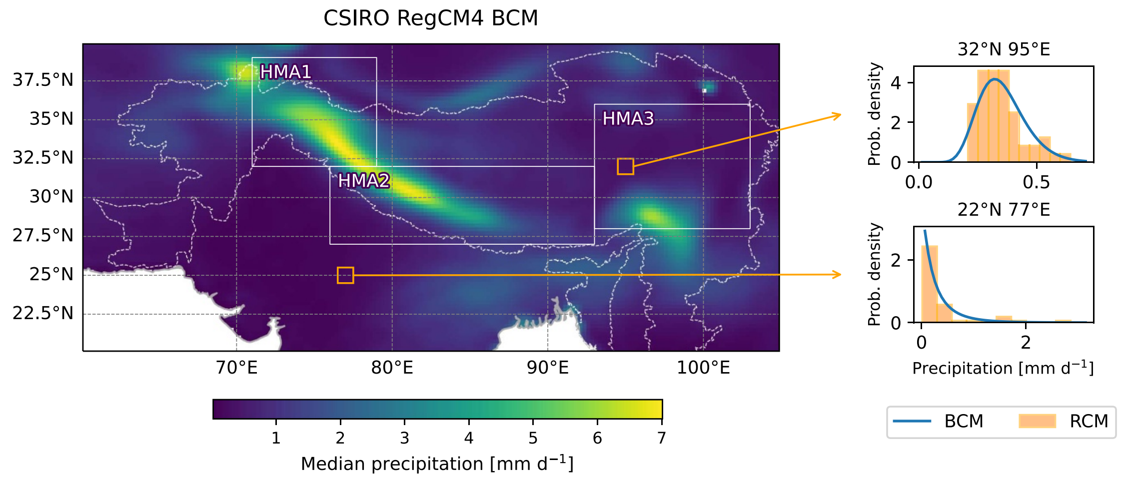

Figure 4. CSIRO RegCM4 BCM distributions between 1951 and 2005 for the month of February. The figure shows the median of the surrogate distributions over HMA (left). The surrogate distributions (blue line) are compared with the empirical histograms (orange bars) for two example locations (top and bottom right). The full surrogate distributions are faithful to the original RCM distributions, which differ significantly depending on time and location.

$$p_{\mathrm{MoE}}\left(f_{\boldsymbol{\lambda}}\left(\boldsymbol{x}_*\right)|\boldsymbol{w},\mathcal{D}\right)=\sum_{r=1}^{R} w_r p_r\left(f_{\lambda_r}\left(\boldsymbol{x}_*\right)|\mathcal{D}_r\right). \quad (9)$$

Here, $\mathcal{D}$ is the dataset for all RCMs, and $\mathcal{D}_r$ is the dataset for the rth RCM for $r=1,\ldots,R$. The distribution $p_r(f_{\lambda_r}(\boldsymbol{x}_*)|\mathcal{D}_r)$ is the *unwarped* rth RCM surrogate distribution conditioned on $\mathcal{D}_r$. It is *not* Gaussian and depends on the warping parameter λ_r The scalar $w_r=w_r(\boldsymbol{x}_*)$ is the weight of the rth RCM at $\boldsymbol{x}_*$. It is constrained to be $\sum_{r=1}^{R} w_r=1$ and $w_r>0$.

3.2.2. Weights

We would like to choose weights $\boldsymbol{w}$ such that the MoE accurately predicts future and past precipitation distributions. Without prior knowledge, the simplest weighting would be $w_r=1/R$, i.e. an equally-weighted mixture of the RCM surrogates, not assigning more importance to any particular RCM. Instead, we inform the weights through a reference dataset (APHRODITE; see section 2.3), having previously downscaled RCM predictions onto the same grid using the surrogate models, as shown in section 3.1.1 and figure 4.

A naive approach would be to directly learn the weights that maximise the likelihood of the parameters (APHRODITE). However, there are two pitfalls to this method. First, the number of effective parameters ($R-1=12$), compared to the number of datapoints at each location and time ($\sim$30), results in overfitting the R model weights (not shown). Second, direct maximum likelihood optimisation could upweigh RCM surrogates that individually predict precipitation poorly, but when summed together best approximate APHRODITE. We want to ensure that RCM surrogates that contribute significantly to the aggregate are independently close fits to APHRODITE. This condition prevents error compensation in the MoE.

Thus, we inform the weights using a measure of distance between the distribution of each RCM surrogate and that of APHRODITE. We choose the Wasserstein distance (Panaretos and Zemel 2019), which originates from optimal transport theory and represents the minimum cost of moving the probability density of one distribution into the shape of another. For two empirical distributions P and Q, the Wasserstein distance W is given by:

$$W(P,Q)=\left(\frac{1}{N}\sum_{i=1}^{N}\|X_i-Y_i\|^p\right)^{1/p}, \quad (10)$$

where X_i and Y_i are the samples of P and Q, respectively, and p is the moment of the distribution we are interested in. Here, we use the first moment $p=1$, which represents the shift of the distribution mean. We calculate the Wasserstein distances with respect to a reference historical period for each APHRODITE grid point and month. Previous research has shown that evaluating the shapes of the distribution rather than the differences in the absolute precipitation outputs better distinguishes how well the models represent precipitation (Martinez-Villalobos and Neelin 2021). We therefore scale the 95th percentile of the BCM outputs to match those of APHRODITE and thus avoid comparing RCMs based on systematic biases in their precipitation output.

We then introduce the statistical temperature T, which interpolates between choosing the most accurate RCM ($T=0$) and equally weighting the RCMs ($T\to\infty$). Using this concept, the weights are parametrised as

$$w_r=\frac{\mathrm{e}^{-h(W_r)/T}}{\sum_{r'}\mathrm{e}^{-h(W_{r'})/T}}, \quad (11)$$

where $h(\cdot)$ is some monotonically increasing function. The statistical temperature T is then optimised to maximise the likelihood on the APHRODITE dataset. We chose $h(\cdot) = \ln(\cdot)$, which reduces the variation between the Wasserstein distances. This choice meets our criteria for the weights to vary smoothly in space and to not choose a single RCM. To further stabilise optimisation, we maximised the likelihood around a 5×5 grid surrounding each location. The RCM surrogates with the smallest Wasserstein distances and the maximum MoE weights over HMA are shown in appendix C.

4. MoE validation

To evaluate the performance of the MoE approach relative to the EW, the historical experiment is divided into two sets: a training period (1951–1980) and a validation period (1981–2005). We use the continuous rank probability score (CRPS, Hersbach 2000) to measure the distance between the predicted (MoE or EW) and observed (APHRODITE) cumulative distribution functions (CDF) F. The CRPS is defined as the integral of the squared difference between the predicted CDF $F(y)$ and the observed CDF $F_{\text{obs}}(y)$ over the range of possible values y:

$$\text{CRPS} = \int_{-\infty}^{\infty} \left(F(y) - F_{\text{obs}}(y)\right)^2 \, \mathrm{d}y. \tag{12}$$

Smaller scores therefore imply more skill.

The spatiotemporal distribution of the differences between the MoE and EW CRPS are presented in figure 5. The MoE CRPS values are in most cases smaller than the EW CRPS values, with the MoE approach yielding an average improvement of 31% over EW. The greatest improvements are over the Karakoram mountains and Himalayas, with an annual average difference of -0.34 mm d^{-1} for the West Himalaya and Karakoram (HMA1), -0.24 mm d^{-1} for the Central and East Himalaya (HMA2), and -0.14 mm d^{-1} for the Hengduan Shan and Southeast Tibet (HMA3).

The differences between MoE and EW can be explained by the relative contribution of RCM members to the ensemble distributions. For example, over certain areas along the central-eastern Tibetan plateau the MoE method assigns dominant weight to one or two RCM surrogates (maximum weight $\geqslant$0.7, figure C2). On the other hand, EW dilutes the skill by not down-weighting members with poorer performance, thus broadening the predictive distribution and inflating CRPS.

In contrast, areas where no significant change in scores is observed tend to correspond to locations where the MoE learns close to equal weighting (see figure C2). This occurs in the south Himalaya during the summer monsoon, where the spread in Wasserstein distance between experts collapses and the learned weights revert towards 1/13. Here, MoE gains largely disappear. Indeed, EW sometimes shows marginally superior performance. Therefore, the spatial pattern of CRPS improvement mirrors the pattern of weight concentration: where model skill heterogeneity is large, MoE expert selection yields significant gains; where expert skill converges, EW averaging suffices. The MoE also improves over taking any individual expert, i.e. RCM surrogate, by 254% as illustrated in figures D1 and D2.

5. Refined climatologies

To generate refined historical and future climatologies, we optimise the weights over the entire historical period (1951–2005), maximising the data available for training. We then use these weights to generate MoE ensemble predictions for the same climate scenarios and time periods as Sanjay *et al* (2017): (i) historical (1976–2005), (ii) RCP4.5 near (2036–2065) and future (2066–2095), and (iii) RCP8.5 near and far future. In the following analysis, we focus on historical and far-future periods to generate projected precipitation trends relative to a historical baseline. The near-future projections show similar trends to the far future, but with less pronounced changes, and are included in appendix F.

5.1. Historical predictions

Figure 6 plots the relative error for median precipitation (REMP) of the MoE and EW with respect to APHRODITE for the historical period. For the monsoon season, the MoE and EW both under- and overestimate precipitation for the Himalaya and Karakoram (HMA1 and HMA2). More specifically, the models predict a wet bias for the Karakoram, inner Himalayas, and western Tibetan Plateau and a dry bias for the outer Himalayas. Precipitation over the eastern Tibetan Plateau, Hengduan Shan, and Southeast Tibet (HMA3) is generally underestimated by both models. For winter, the MoE again under- and overestimates precipitation across the three studied subregions, but with larger wet biases, while the EW largely overestimates precipitation across all three areas during winter.

The third row of figure 6 plots the difference between the MoE and EW REMP. The MoE makes large improvements over the EW, particularly in locations where the EW overestimates precipitation. This corresponds to a difference in REMP of 2.02 for the West Himalaya and Karakoram (HMA1), 0.32 for the Central and East Himalaya (HMA2), and 0.08 for the Hengduan Shan and Southeast Tibet (HMA3) over the summer. During winter, larger improvements are obtained with REMP differences of 1.58, 3.02, and 2.33 for the West Himalaya and Karakoram (HMA1), the Central and East Himalaya (HMA2), and the Hengduan Shan and Southeast Tibet (HMA3), respectively. In summary, the MoE predicts smaller median precipitation values compared to the EW, more in line with APHRODITE.

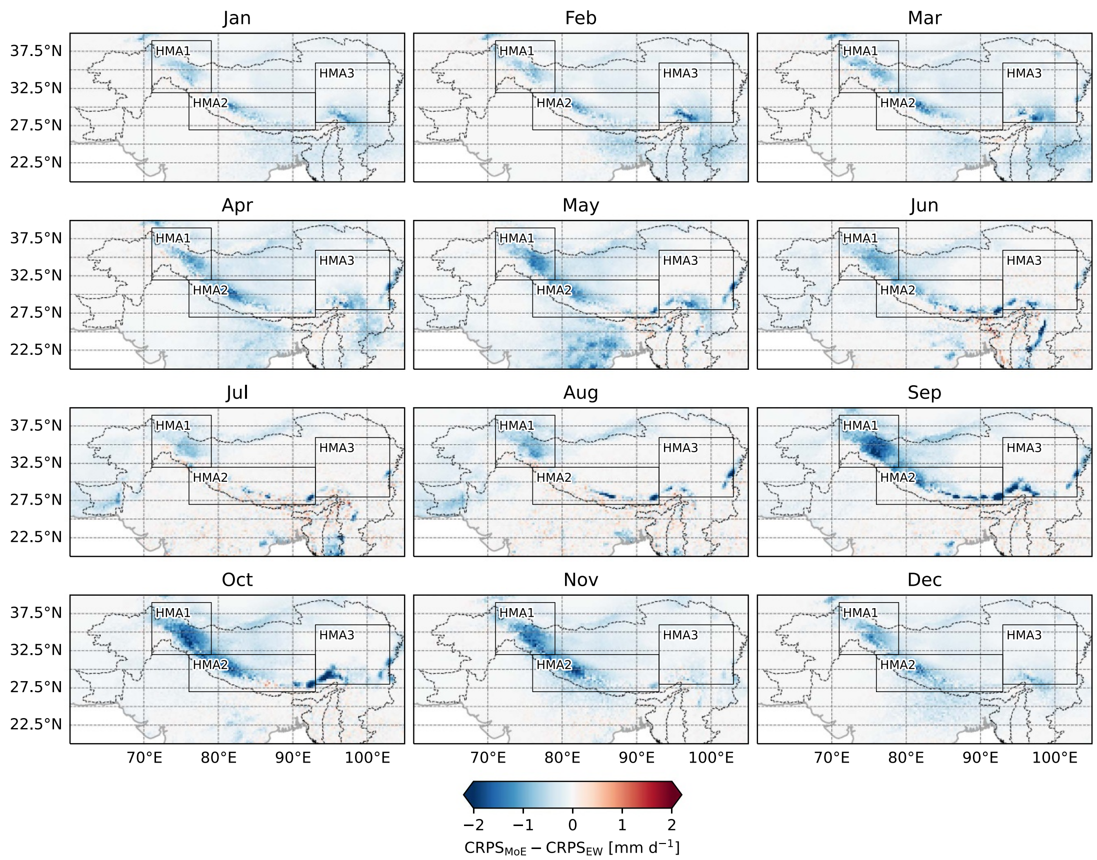

Figure 5. MoE and EW CRPS differences over the held-out validation period (1981–2005). The MoE and EW outputs are generated from 10^5 RCM surrogate samples for each month and location. Negative values (blue) imply MoE matches APHRODITE more closely while positive values (red) represent times and locations where the EW performs better. The MoE outperforms the EW for each month, in particular over the Karakoram and Himalayan arc.

However, MoE predictions are still subject to large errors, and bias correction would still be needed before use. For completeness, we include the historical predictions biases in mm d^{-1} in appendix E.

5.2. Future projections

Figure 7 shows far-future MoE projections with respect to the historical period. For RCP4.5 and RCP8.5, the monsoon season is projected to see an increase in median precipitation over most of HMA, with the exception of the north Hindu Kush mountains. The greatest changes occur over Kunlun mountains and Southeast Tibet. More specifically, the MoE predicts an average increase of 15% for the West Himalaya and Karakoram (HMA1), 23% for the Central and East Himalaya (HMA2), and 27% for the Hengduan Shan and Southeast Tibet (HMA3) for RCP8.5. Compared to RCP4.5, RCP8.5 presents a more pronounced wetting trend over the study area.

For winter, the MoE projects an increase in median precipitation over the north and north-east HKH, with the largest changes occurring over the Quilian Shan and East Kunlun mountains. The Hindu Kush mountains and the outer Himalayas show an overall decrease in precipitation for this season. These trends lead to mixed predictions for the West Himalaya and Karakoram (HMA1) and the Central and East Himalaya (HMA2), with average relative changes of 15% and −7% for RCP8.5, respectively. Winter precipitation over the Hengduan Shan and Southeast Tibet (HMA3) is expected to increase with an average relative change of 27%. Compared to RCP4.5, RCP8.5 generally presents more pronounced wetting and drying trends during winter.

Finally, the relative changes in precipitation from the MoE can be compared with those from the EW. Figure 8 plots the difference between the MoE and EW relative changes for the far-future RCP8.5 scenario with respect to their historical reference predictions. The plot shows that for large areas, including the Tibetan Plateau over summer and the Hindu Kush, the relative precipitation changes predicted by the MoE and the EW stay within ±10% across different percentiles. However, many locations also present significant changes.

During the summer monsoon, positive differences in predicted precipitation changes (i.e. MoE increase relative to EW) are observed over the West Himalaya, Karakoram and Kunlun mountains with median and 95th percentile differences reaching maximum values of 54% and 32% for HMA1. This suggests that the probability of events associated with

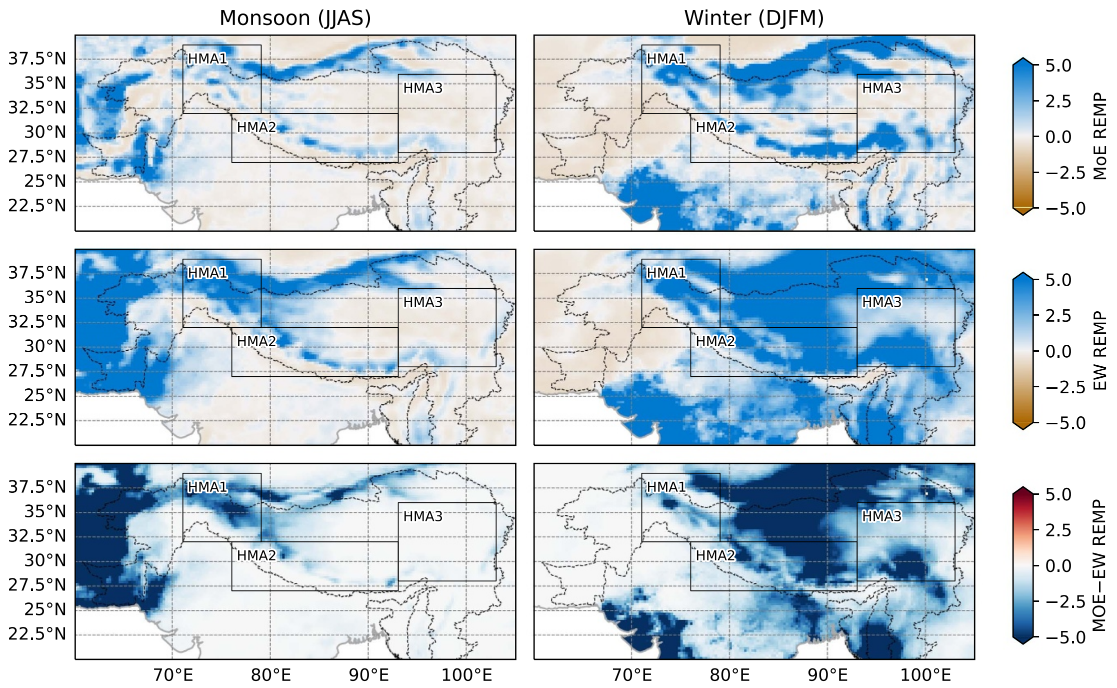

Figure 6. Historical MoE and EW relative error for median precipitation (REMP) with respect to APHRODITE over HMA. The MoE REMP (top), the EW REMP (middle), and the REMP difference between MoE and EW (bottom) are plotted for the historical reference period (1976–2005) for the summer monsoon (left) and winter (right). The MoE and EW outputs are generated from 10^5 RCM surrogate samples for each month and location. The MoE makes large improvements over the EW, in particular, over locations where precipitation is overestimated by the EW.

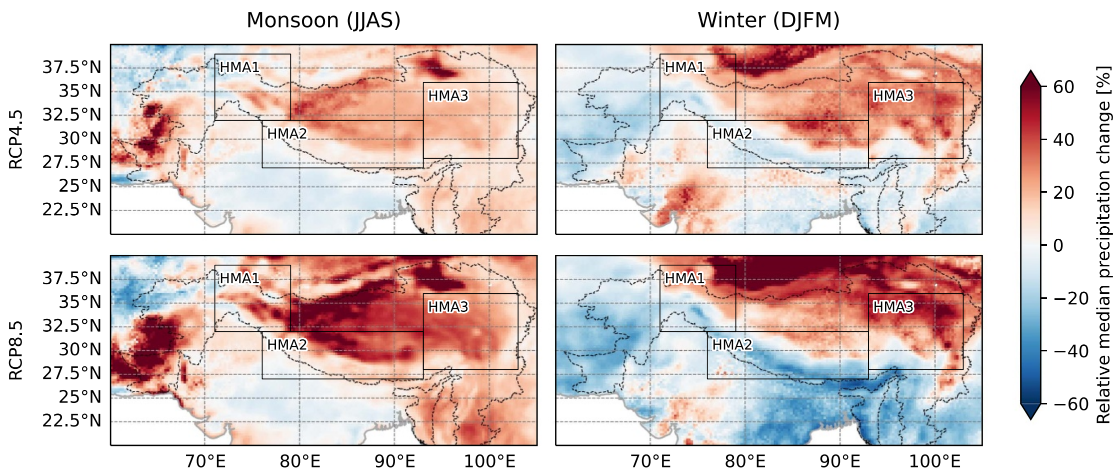

Figure 7. Relative changes between far-future (2066–2095) and historical (1976–2005) MoE precipitation. The plot shows the median changes for RCP4.5 (top) and RCP8.5 (bottom) during the summer monsoon (left) and winter (right) for the far-future across HMA. The MoE output is generated from 10^5 RCM surrogate samples for each month and location. The MoE predicts large departures from the precipitation predicted for the historical reference period.

very high precipitation rates, such as floods and landslides, could be higher compared to previous estimates. During winter, the MoE predicts a greater decrease in projected precipitation over the inner Himalayas and Karakoram, with median differences achieving minimum values of −31% for HMA1 and −27% for HMA2. This could signify a decrease in the contribution of winter precipitation to solid water resources. At the same time, large positive differences are observed over the Tibetan Plateau and Southeast Tibet, with maximum differences in median values of 51% for HMA2 and 62% for HMA3.

6. Discussion

This paper explored probabilistic ensemble learning to aggregate RCMs. Using a MoE model, precipitation predictions were improved compared to an

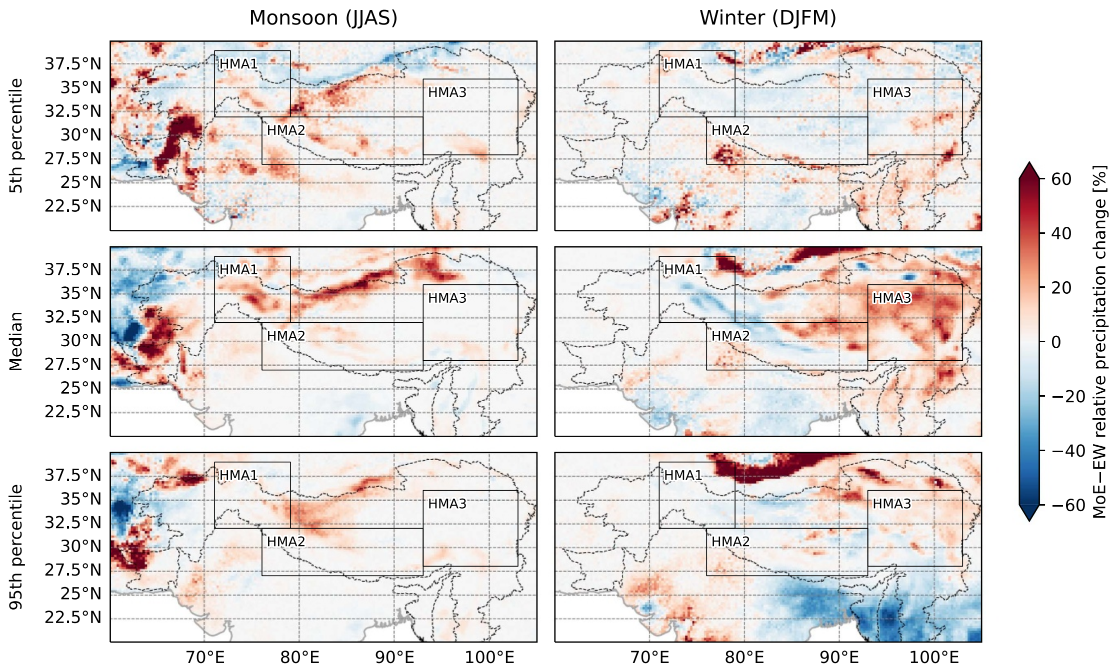

Figure 8. MoE and EW relative prediction differences for the far future (2066–2095) under RCP8.5 across HMA with respect to their historical reference predictions. The plot shows the differences between the predicted relative changes for the 5th percentile (bottom), median (middle), and 95th percentile (top) during the summer monsoon (left) and winter (right). The MoE and EW outputs are generated from 10^5 RCM surrogate samples for each month and location. Although for many locations the MoE and EW projections are similar, there are significant difference in predictions during both the summer monsoon and winter over key mountain ranges.

EW approach. Unlike other ensembling techniques, this method does not narrow the non-discountable envelope of climate change, but refines it. With GPs as its foundational building blocks, the MoE should work equally well with RCMs on different grids and can be straightforwardly applied to any kind of model ensemble. However, this study does present several limitations.

First, we assume that the driving variables from the driving GCMs are sufficiently well-modelled to dynamically downscale precipitation. In reality, CMIP5 GCMs present large differences in their predictions over HMA (Palazzi *et al* 2015, Panday *et al* 2015). Moreover, recent studies incorporating more realistic aerosol scenarios further amplify future precipitation uncertainty (Wilcox *et al* 2020, Jiang *et al* 2023). This poor representation of precipitation in RCMs can be observed in this study through the optimal weights chosen by the MoE: when no model in the ensemble performs particularly well, the MoE chooses weights close to the EW. This is the case for the Hindu Kush during winter and for the Tibetan Plateau during summer (cf figures 6 and C2). This behaviour, in turn, implies poor representation of the two main precipitation drivers for HMA: the East Asian summer monsoon and western disturbances. Further work could use this framework to explain performance differences across RCMs by analysing other output variables, such as pressure or humidity fields.

Second, the methods used to create the RCM surrogates have several limitations. By applying GPs, we assume that the RCM outputs are unimodal. In reality, this is not always the case, and we could be missing out on inter-annual variability. The degree to which the individual GP samples are consistent in time and space is also uncertain. Using a probabilistic surrogate for the RCM means that future precipitation scenarios can be sampled thanks to the correlations learnt between different locations and times. However, by aggregating data over time, the RCM surrogate, and by extension the MoE, could plausibly sample two high precipitation values for neighbouring locations or months, although the joint event is statistically unlikely. Reintroducing some of the information from the RCM time series would be needed to ensure further consistency.

Third, the weighting function used to aggregate the RCMs, although well motivated, has an arbitrary form. There is significant room to refine the approach through more case studies. Finally, monthly precipitation is not the only variable linked to water security over HMA. To get a more holistic view of future changes, and thus the risk of extreme events such as floods and droughts, the experiments in this paper could be repeated with variables such as temperature or the number of cumulative dry and wet days. Extreme value analysis could also be applied to calculate updated rates of return for single or compounding hydrological extremes.

7. Conclusion

This paper investigated a MoE approach to aggregate an ensemble of RCMs. This ensemble learning method is generalisable to other climate variables and model ensembles and could be applied to model aggregation problems in other fields. The MoE was applied to the CORDEX-WAS ensemble over HMA and was found to be more accurate than the EW average or any single ensemble member for a held-out validation period. Compared to the EW, the MoE projects wetter summers but drier winters over the western Himalayas and Karakoram and wetter winters over the Tibetan Plateau, Hengduan Shan, and South East Tibet for both the near and far future under RCP4.5 and RCP8.5.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.14907030 (2025); https://doi.org/10.5281/zenodo.14837272 (2025); https://esgf-metagrid.cloud.dkrz.de/; http://aphrodite.st.hirosaki-u.ac.jp/download/.

Acknowledgment

The authors thank Andrew Orr and Scott J Hosking for their thoughtful feedback on the manuscript. This work was supported by the Engineering and Physical Sciences Research Council through the AI for Environmental Risk (AI4ER) CDT (Grant Number: EP/S022961/1) for K T and M G M, and the DTP International Studentship (Grant Number: EP/W524475/1) for SWPK.

Appendix A. BCM softmax-variance control

Following Cohen *et al* (2020), we use their general expression for expert weights:

$$\beta_j(\boldsymbol{x}_*) \propto \exp\left(-1/\psi_j(\boldsymbol{x}_*)\,\tau\right), \quad \sum_{j=1}^{M} \beta_j(\boldsymbol{x}_*) = 1, \tag{A.1}$$

where τ is a temperature parameter (with a similar function to the variable T in section 3.2) that controls the sparsity between experts by multiplicatively compounding the weights of stronger experts. The function $\psi_j(\boldsymbol{x}_*)$ describes the level of confidence of the jth expert at test point $\boldsymbol{x}_*$. Like Cohen *et al* (2020), we set $\psi_j(\boldsymbol{x}_*)$ to the posterior predictive variance at $\boldsymbol{x}_*$, i.e. $\sigma_j^2(\boldsymbol{x}_*)$ and choose $\tau = 1/8$.

Appendix B. MoE probability distribution

The transformed precipitation $\tilde{y}$ for the rth RCM surrogate at some input coordinate $\boldsymbol{x}_*$ is assumed to be normally distributed. Suppressing input variables and fixed hyperparameters, we have

$$p_r(\tilde{y}_r) = \frac{1}{\sqrt{2\pi\,\sigma_r^2}} \exp\left(-\frac{(\tilde{y}_r - \mu_r)^2}{2\sigma_r^2}\right). \tag{B.1}$$

The distribution for y can then be found by transforming the coordinates and calculating the Jacobian $p_r(\tilde{y}_r(y))\frac{\mathrm{d}\tilde{y}_r}{\mathrm{d}y}$, yielding

$$p_r'(y) = \frac{y^{\lambda_r - 1}}{\sqrt{2\pi\,\sigma_r^2}} \exp\left(-\frac{\left(\frac{y^{\lambda_r}-1}{\lambda_r} - \mu_r\right)^2}{2\sigma_r^2}\right). \tag{B.2}$$

However, there is a small issue with this procedure. Inverting the transformation, we find that for $\lambda_r \neq 0$, $y(\tilde{y}) = (\lambda_r\tilde{y} + 1)^{1/\lambda_r}$. This quantity is ambiguous for $\tilde{y}_r < -1/\lambda_r$, as the fractional power of a negative number is ill-defined. To remedy this, we restrict the domain of $\tilde{y}$ to $(-1/\lambda_r, \infty)$, with the cost that we need to renormalise the probability distribution by its integral. Defining

$$C_r := \int_{-1/\lambda_r}^{\infty} \mathrm{d}\tilde{y}_r p_r'(\tilde{y}_r) = \int_{-(\lambda_r^{-1}+\mu_r)/\sigma_r}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}$$
$$= \frac{1}{2}\left(1 + \mathrm{erf}\left[\frac{1}{\sqrt{2\sigma_r^2}}\left(\frac{1}{\lambda_r} + \mu_r\right)\right]\right), \tag{B.3}$$

the properly normalised probability distribution for y is given by

$$p_r(y) = \frac{y^{\lambda_r - 1}}{C_r\sqrt{2\pi\,\sigma_r^2}} \exp\left(-\frac{\left(\frac{y^{\lambda_r}-1}{\lambda_r} - \mu_r\right)^2}{2\sigma_r^2}\right), \tag{B.4}$$

with domain $y \in [0, \infty)$. In practice, we expect C_r to be very close to one.

Finally, given some weights $\{w_r\}_r$, the MoE's distribution is given by

$$p(y) = \sum_r w_r p_r(y), \tag{B.5}$$

with the constraint that $\sum_r w_r = 1$.

Appendix C. Which RCM, when and where?

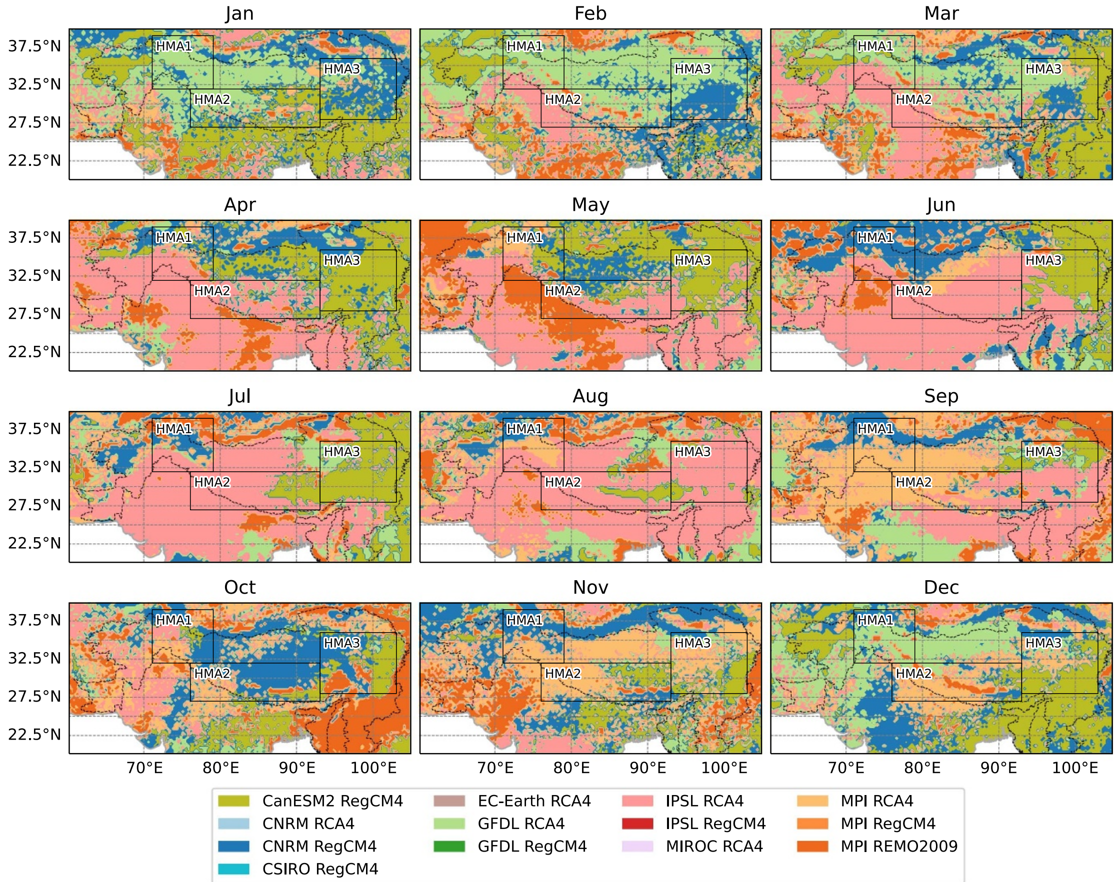

Figure C1. RCMs with the smallest Wasserstein distances for each month and location over HMA. The distances are calculated using the scaled distributions of precipitation between 1951 and 2005. The spatial distribution of these models is not random but follows distinct spatiotemporal patterns.

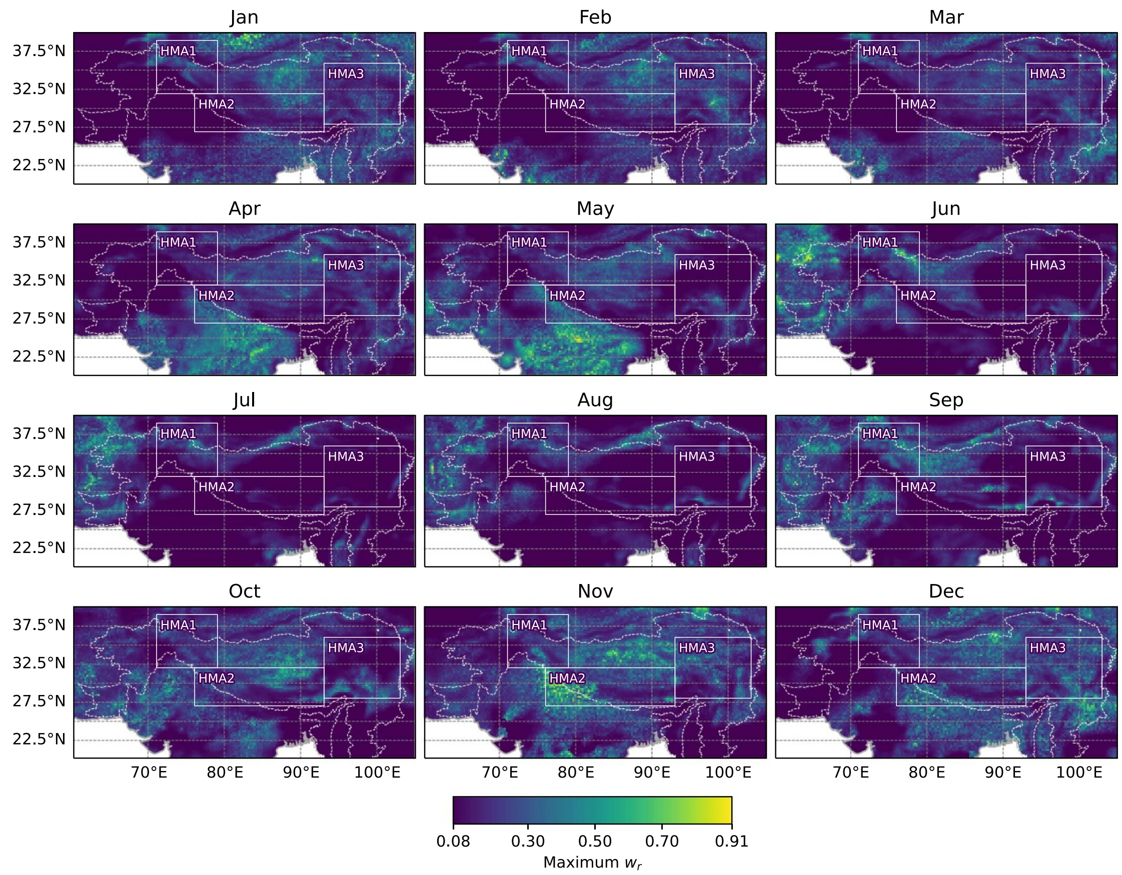

Figure C2. Maximum MoE weights for each month and location over HMA. The weights are optimised between 1951 and 2005. Here, one RCM surrogate (BCM) is never chosen over all others, i.e. the maximum w_r is never 1. The maximum BCM weighting is highest over the Hindu Kush during the summer and over the Tibetan Plateau during the winter. Conversely, the MoE with close to equal-weights (i.e. maximum $w_r \rightarrow 1/13$) is more advantageous during the summer monsoon over the Tibetan Plateau and during the winter over Hindu Kush.

Appendix D. MoE and BCM CRPS differences

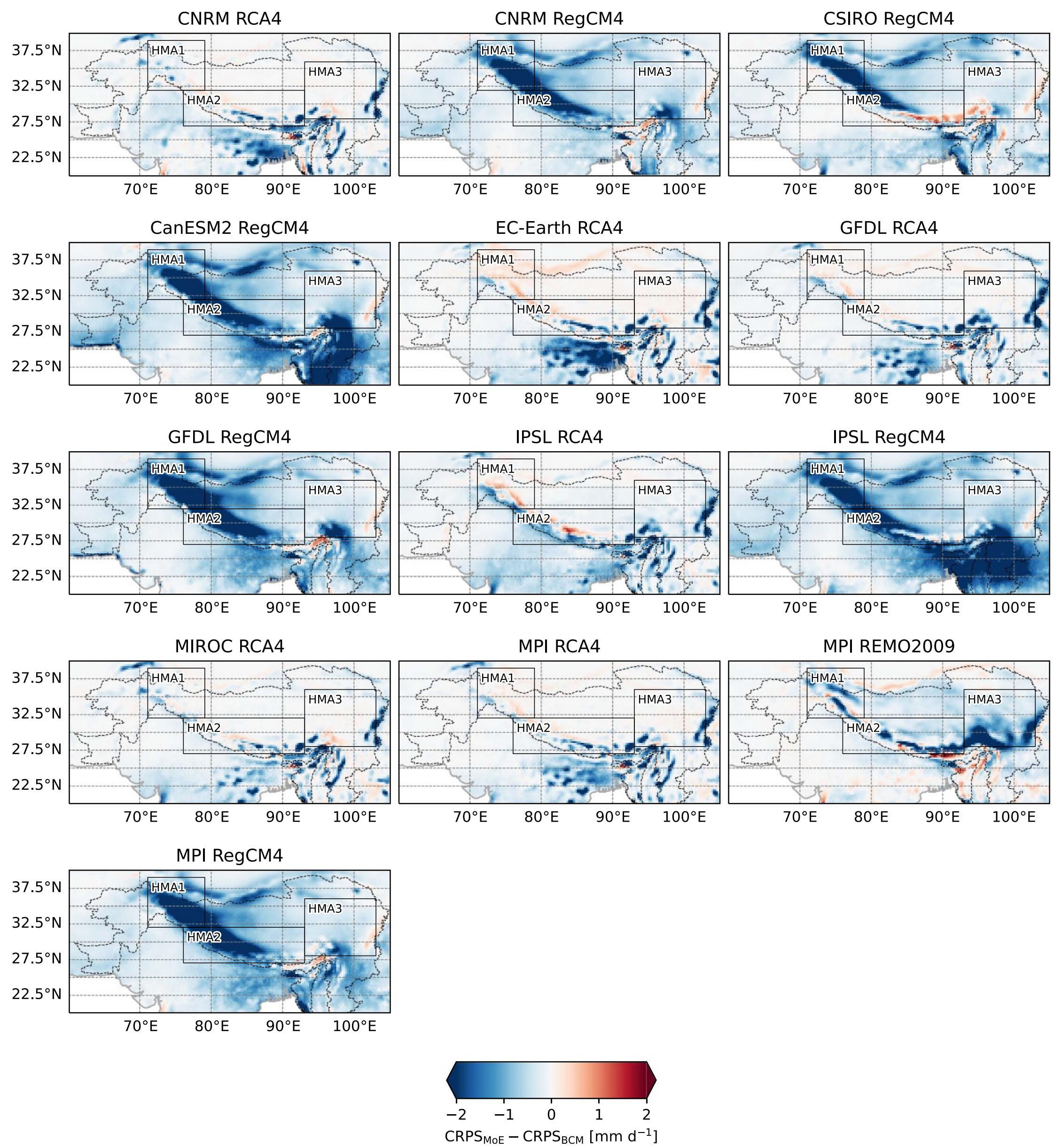

Figure D1. Annual CRPS difference between the MoE and RCM surrogates (BCM) for the held-out validation period (1981–2005). The MoE outputs are generated from 10^5 RCM surrogate samples for each month and location. Negative values (blue) imply MoE matches APHRODITE more closely while positive values (red) represent times and locations where the BCM performs better. Overall, the MoE outperforms the BCMs over the entire year with large improvement over HMA1 and HMA2, especially for the RegCM4 surrogates.

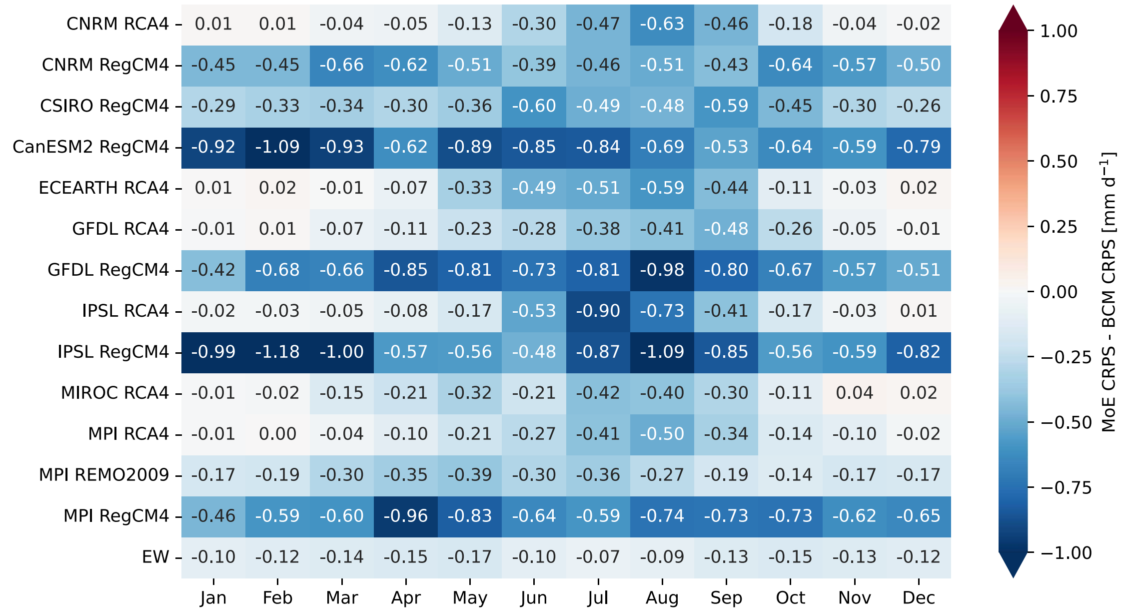

Figure D2. Scorecard showing the differences between the CRPS yielded by the MoE and that of each RCM surrogate (BCM) as well as the EW over the held-out validation period (1981–2005). Each cell represents the difference for a specific model and month, averaged over the entire spatial domain. The MoE and EW outputs are generated from 10^5 RCM surrogate samples for each month and location. Positive values (red) indicate higher CRPS for the MoE compared to the BCM, while negative values (blue) indicate lower CRPS for the MoE compared to the BCM.

Appendix E. Historical predictions biases

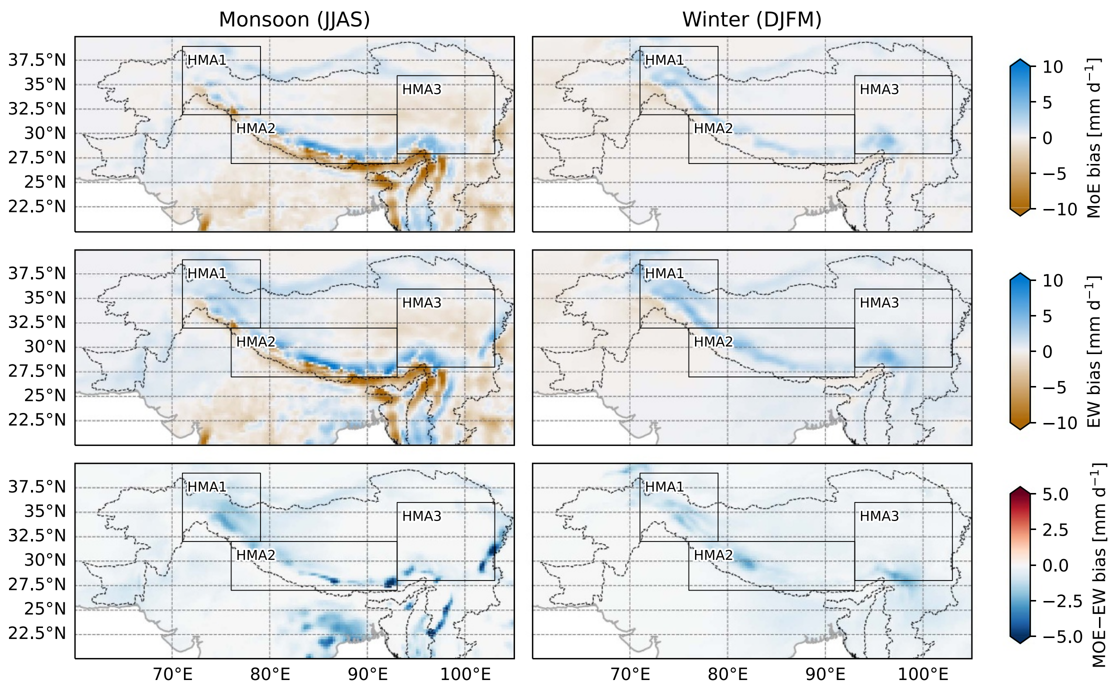

Figure E1. Historical MoE and EW bias for median precipitation with respect to APHRODITE over HMA. The MoE bias (top), the EW bias (middle), and the bias difference between MoE and EW (bottom) are plotted for the historical reference period (1976–2005) for the summer monsoon (left) and winter (right). The MoE and EW outputs are generated from 10^5 RCM surrogate samples for each month and location. The MoE makes large improvements over the EW, in particular, over locations where precipitation is overestimated by the EW.

Appendix F. Near-future predictions

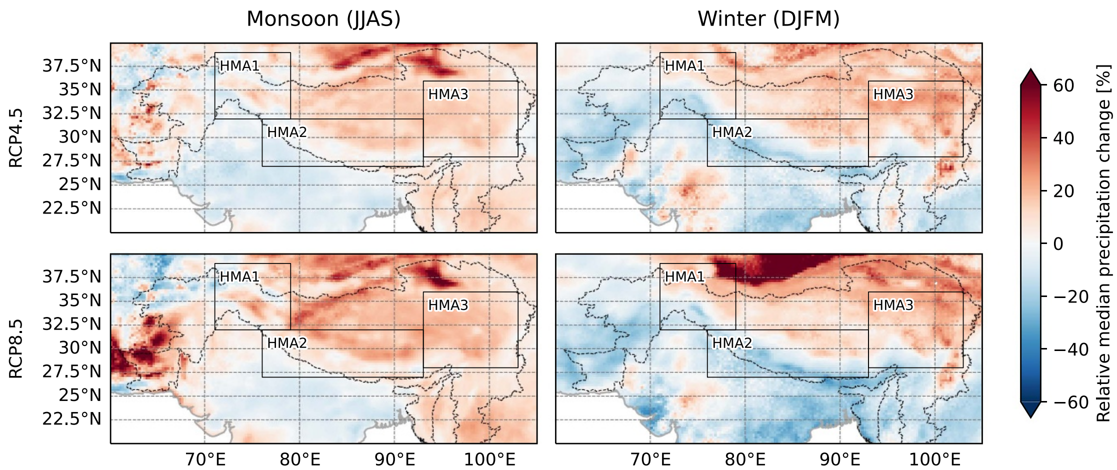

Figure F1. Relative changes between near-future (2036–2065) and historical (1976–2005) MoE precipitation across HMA. The plot shows the median changes for RCP4.5 (top) and RCP8.5 (bottom) during the summer monsoon (left) and winter (right). The MoE and EW outputs are generated from 10^5 RCM surrogate samples for each month and location. Results are similar to far-future predictions with less pronounced changes.

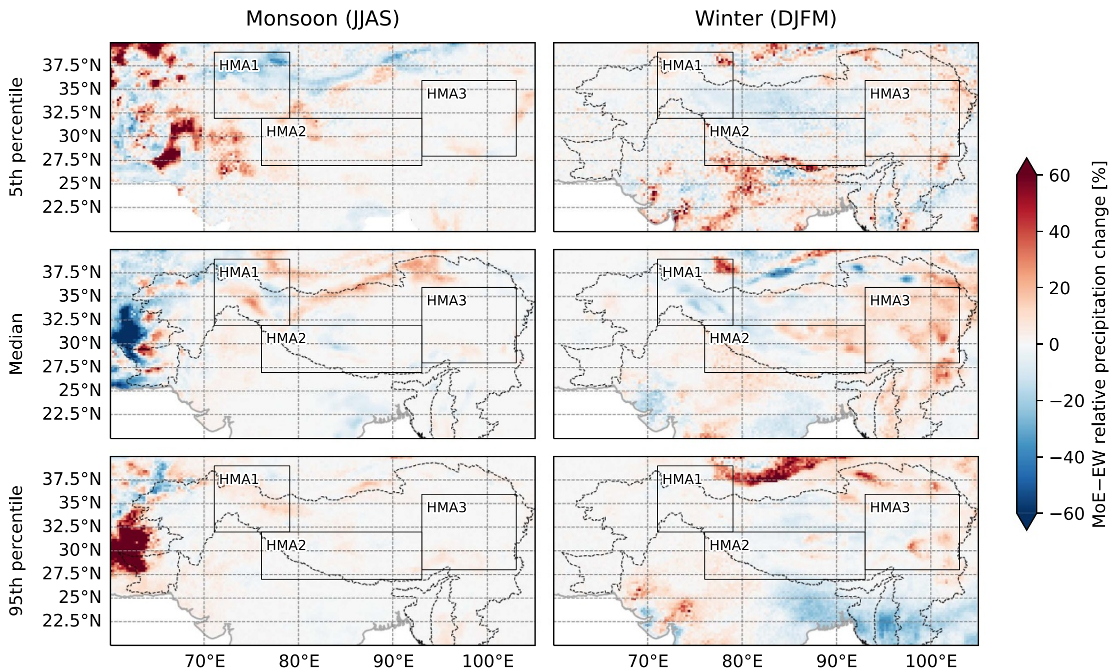

Figure F2. MoE and EW relative prediction differences for the near future (2066–2095) under RCP8.5 across HMA with respect to their historical reference predictions. The plot shows the difference between the predicted relative changes for the 5th percentile (bottom), median (middle), and 95th percentile(top) during the summer monsoon (left) and winter (right). The MoE and EW outputs are generated from 10^5 RCM surrogate samples for each month and location. Results are similar to far-future predictions with less pronounced differences.

References

Amos M, Pinder T and Young P 2022 Identifying latent climate signals using sparse hierarchical Gaussian processes *NeurIPS 2022 Workshop on Tackling Climate Change With Machine Learning* (available at: www.climatechange.ai/papers/neurips2022/15) (Accessed 18 November 2022)

Bajracharya S R and Shrestha B R 2011 The status of glaciers in the Hindu Kush-Himalayan region *Int. Centre for Integrated Mountain Development (ICIMOD)* (available at: https://doi.org/10.53055/ICIMOD.551)

Box G E and Cox D R 1964 An analysis of transformations *J. R. Stat. Soc. B* **26** 211–43

Chylek P, Li J, Dubey M, Wang M and Lesins G 2011 Observed and model simulated 20th century Arctic temperature variability: Canadian earth system model CanESM2 *Atmos. Chem. Phys. Discuss.* **11** 22893–907

Cohen S, Mbuvha R, Marwala T and Deisenroth M 2020 Healing products of Gaussian process experts *Proc. 37th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research*) vol 119 (PMLR) pp 2068–77 (available at: https://proceedings.mlr.press/v119/cohen20b.html)

Deisenroth M and Ng J W 2015 Distributed Gaussian processes *Proc. 32nd Int. Conf. on Machine Learning* (*Proc. Machine Learning Research*) (*Lille, France*, *07 –09 July 2015*) vol 37 (PMLR) pp 1481–90 (available at: https://proceedings.mlr.press/v37/deisenroth15.html)

Dimri A 2021 Bias correction demonstration in two of the Indian Himalayan river basins *J. Water Clim. Change* **12** 1297–309

Dufresne J-L *et al* 2013 Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5 *Clim. Dyn.* **40** 2123–65

Dunne J P *et al* 2012 GFDL's ESM2 global coupled climate–carbon earth system models. part I: physical formulation and baseline simulation characteristics *J. Clim.* **25** 6646–65

Eyring V *et al* 2019 Taking climate model evaluation to the next level *Nat. Clim. Change* **9** 102–10

Flato G *et al* 2014 Evaluation of climate models *Climate Change 2013: the Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press) pp 741–866

Giorgi F *et al* 2012 RegCM4: model description and preliminary tests over multiple CORDEX domains *Clim. Res.* **52** 7–29

Giorgi F and Gutowski W J 2015 Regional dynamical downscaling and the CORDEX initiative *Ann. Rev. Environ. Resour.* **40** 467–90

Girona-Mata M *et al* 2024 Probabilistic precipitation downscaling for ungauged mountain sites: a pilot study for the Hindu Kush Karakoram Himalaya *EGUsphere* **2024** 1–33

GTN-G 2023 GTN-G glacier regions [Data set] *Global Terrestrial Network for Glaciers* (available at: https://doi.org/10.5904/gtng-glacreg-2023-07)

Harris G R, Sexton D M, Booth B B, Collins M and Murphy J M 2013 Probabilistic projections of transient climate change *Clim. Dyn.* **40** 2937–72

Hausfather Z, Marvel K, Schmidt G A, Nielsen-Gammon J W and Zelinka M 2022 Climate simulations: Recognize the 'hot model' problem *Nature* **605** 26–29

Hazeleger W *et al* 2010 EC-Earth: a seamless earth-system prediction approach in action *Bull. Am. Meteorol. Soc.* **91** 1357–64

Hersbach H 2000 Decomposition of the continuous ranked probability score for ensemble prediction systems *Weather Forecast.* **15** 559–70

ICIMOD 2008 Outline boundary of Hindu Kush Himalayan (HKH) region [data set] International Centre for Integrated Mountain Development (ICIMOD)

Immerzeel W W *et al* 2020 Importance and vulnerability of the world's water towers *Nature* **577** 364–9

Jeffrey S, Rotstayn L, Collier M, Dravitzki S, Hamalainen C, Moeseneder C, Wong K and Syktus J 2013 Australia's CMIP5 submission using the CSIRO-Mk3.6 model *Aust. Meteorol. Oceanogr. J.* **63** 1–13

Jiang J, Zhou T, Qian Y, Li C, Song F, Li H, Chen X, Zhang W and Chen Z 2023 Precipitation regime changes in High Mountain Asia driven by cleaner air *Nature* **623** 544–9

Jungclaus J H, Fischer N, Haak H, Lohmann K, Marotzke J, Matei D, Mikolajewicz U, Notz D and Storch J V 2013 Characteristics of the ocean simulations in the Max Planck Institute Ocean Model (MPIOM) the ocean component of the MPI-Earth system model *J. Adv. Model. Earth Syst.* **5** 422–46

Kharin V V and Zwiers F W 2002 Climate predictions with multimodel ensembles *J. Clim.* **15** 793–9

Knutti R 2010 The end of model democracy? an editorial comment *Clim. Change* **102** 395–404

Lee J-Y *et al* 2021 Future global climate: scenario-based projections and near-term information *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press) pp 553–672

Lehner B and Grill G 2013 Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems *Hydrol. Process.* **27** 2171–86

Martinez-Villalobos C and Neelin J D 2021 Climate models capture key features of extreme precipitation probabilities across regions *Environ. Res. Lett.* **16** 024017

Masson D and Knutti R 2011 Climate model genealogy *Geophys. Res. Lett.* **38** 6864

Massoud E C, Lee H K, Terando A and Wehner M 2023 Bayesian weighting of climate models based on climate sensitivity *Commun. Earth Environ.* **4** 365

Maussion F, Scherer D, Mölg T, Collier E, Curio J and Finkelnburg R 2014 Precipitation seasonality and variability over the Tibetan Plateau as resolved by the High Asia Reanalysis *J. Clim.* **27** 1910–27

Min S-K, Simonis D and Hense A 2007 Probabilistic climate change predictions applying Bayesian model averaging *Phil. Trans. R. Soc. A* **365** 2103–16

Nie Y *et al* 2021 Glacial change and hydrological implications in the Himalaya and Karakoram *Nat. Rev. Earth Environ.* **2** 91–106

Norris J, Carvalho L M, Jones C and Cannon F 2019 Deciphering the contrasting climatic trends between the central Himalaya and Karakoram with 36 years of WRF simulations *Clim. Dyn.* **52** 159–80

Norris J, Carvalho L M, Jones C and Cannon F 2020 Warming and drying over the central Himalaya caused by an amplification of local mountain circulation *npj Clim. Atmos. Sci.* **3** 1

Norris J, Carvalho L M, Jones C, Cannon F, Bookhagen B, Palazzi E and Tahir A A 2017 The spatiotemporal variability of precipitation over the Himalaya: evaluation of one-year WRF model simulation *Clim. Dyn.* **49** 2179–204

Olson R, Fan Y and Evans J P 2016 A simple method for Bayesian model averaging of regional climate model projections: application to southeast Australian temperatures *Geophys. Res. Lett.* **43** 7661–9

Orr A *et al* 2022 Knowledge priorities on climate change and water in the Upper Indus Basin: a horizon scanning exercise to identify the top 100 research questions in social and natural sciences *Earth's Future* **10** e2021EF002619

Orr A, Listowski C, Couttet M, Collier E, Immerzeel W, Deb P and Bannister D 2017 Sensitivity of simulated summer monsoonal precipitation in Langtang Valley, Himalaya, to cloud microphysics schemes in WRF *J. Geophys. Res. Atmos.* **122** 6298–318

Palazzi E, von Hardenberg J, Terzago S and Provenzale A 2015 Precipitation in the Karakoram-Himalaya: a CMIP5 view *Clim. Dyn.* **45** 21–45

Palmer T, Doblas-Reyes F, Hagedorn R and Weisheimer A 2005 Probabilistic prediction of climate using multi-model ensembles: from basics to applications *Phil. Trans. R. Soc. B* **360** 1991–8

Panaretos V M and Zemel Y 2019 Statistical aspects of Wasserstein distances *Ann. Rev. Stat. Appl.* **6** 405–31

Panday P K, Thibeault J and Frey K E 2015 Changing temperature and precipitation extremes in the Hindu Kush-Himalayan region: an analysis of CMIP3 and CMIP5 simulations and projections *Int. J. Climatol.* **35** 3058–77

Rasmussen C E and Williams C K 2006 *Gaussian Processes for Machine Learning* vol 1 (MIT Press)

RGI Consortium 2017 Randolph Glacier Inventory–a dataset of global glacier outlines: Version 6.0 [data set] National Snow and Ice Data Center (available at: https://doi.org/10.7265/4m1f-gd79)

Samuelsson P, Gollvik S, Kupiainen M, Kourzeneva E and van de Berg W J 2015 The surface processes of the Rossby Centre regional atmospheric climate model (RCA4) *Technical Report* SMHI

Sanderson B M and Knutti R 2012 On the interpretation of constrained climate model ensembles *Geophys. Res. Lett.* **39** 2665

Sanjay Jayanarayanan Scientist 2025 Centre for climate change research at indian institute of tropical meteorology *Personal correspondence*

Sanjay J, Krishnan R, Shrestha A B, Rajbhandari R and Ren G-Y 2017 Downscaled climate change projections for the Hindu Kush Himalayan region using CORDEX South Asia regional climate models *Adv. Clim. Change Res.* **8** 185–98

Saul A D, Hensman J, Vehtari A and Lawrence N D 2016 Chained Gaussian processes *Proc. 19th Int. Conf. on Artificial Intelligence and Statistics (Cadiz, Spain, 09–11 May 2016)* (PMLR) pp 1431–40

Snelson E, Ghahramani Z and Rasmussen C 2003 Warped Gaussian processes *Advances in Neural Information Processing Systems (Vacouver, BC, Canada, 8–13 December 2003)* vol 16

Stainforth D A, Downing T E, Washington R, Lopez A and New M 2007 Issues in the interpretation of climate model ensembles to inform decisions *Phil. Trans. R. Soc. A* **365** 2163–77

Taylor K E, Stouffer R J and Meehl G A 2012 An overview of CMIP5 and the experiment design *Bull. Am. Meteorol. Soc.* **93** 485–98

Tazi K 2025 Data for 'Refined climatologies of future precipitation over High Mountain Asia using probabilistic ensemble learning' (0.0.0) [Data set] *Zenodo* (https://doi.org/10.5281/zenodo.14837272)

Tazi K, Girona-Mata M and Kim S W 2025 kenzaxtazi/bcm4rcm: First release (v1.0.0) *Zenodo* (https://doi.org/10.5281/zenodo.14907030)

Tazi K, Orr A O, Hernandez-González J, Hosking S and Turner R E 2024 Downscaling precipitation over High-mountain Asia using multi-fidelity Gaussian processes: improved estimates from ERA5 *Hydrol. Earth Syst. Sci.* **28** 4903–25

Tazi K, Orr A O, Hosking S and Turner R E 2025 Precipitation prediction from large-scale climatic features over the Upper Indus Basin using non-stationary Gaussian processes *Environ. Data Sci.* accepted

Tebaldi C and Knutti R 2007 The use of the multi-model ensemble in probabilistic climate projections *Phil. Trans. R. Soc. A* **365** 2053–75

Teichmann C *et al* 2013 How does a regional climate model modify the projected climate change signal of the driving GCM: a study over different CORDEX regions using remo *Atmosphere* **4** 214–36

Tél T, Bódai T, Drótos G, Haszpra T, Herein M, Kaszás B and Vincze M 2020 The theory of parallel climate realizations: a new framework of ensemble methods in a changing climate: an overview *J. Stat. Phys.* **179** 1496–530

Titsias M 2009 Variational learning of inducing variables in sparse Gaussian processes *Proc. 12th Int. Conf. on Artificial Intelligence and Statistics (Clearwater Beach, FL, USA, 16–18 April 2009)* vol 5 (PMLR) pp 567–74

Tresp V 2000 A Bayesian committee machine *Neural Comput.* **12** 2719–41

Voldoire A *et al* 2013 The CNRM-CM5.1 global climate model: description and basic evaluation *Clim. Dyn.* **40** 2091–121

Watanabe M *et al* 2010 Improved climate simulation by MIROC5: mean states, variability and climate sensitivity *J. Clim.* **23** 6312–35

Wester P, Chaudhary S, Chettri N, Maharjan A, Nepal S and Steiner J 2023 *Water, ice, Society and Ecosystems in the Hindu Kush Himalaya: an Outlook* (International Centre for Integrated Mountain Development)

Westermann S *et al* 2024 ESA Permafrost Climate Change Initiative: Permafrost version 4 data products [data set] NERC EDS Centre for Environmental Data Analysis

Wilcox L J *et al* 2020 Accelerated increases in global and Asian summer monsoon precipitation from future aerosol reductions *Atmos. Chem. Phys.* **20** 11955–77

Yatagai A, Kamiguchi K, Arakawa O, Hamada A, Yasutomi N and Kitoh A 2012 APHRODITE: constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges *Bull. Am. Meteorol. Soc.* **93** 1401–15

Zelinka M D, Myers T A, McCoy D T, Po-Chedley S, Caldwell P M, Ceppi P, Klein S A and Taylor K E 2020 Causes of higher climate sensitivity in CMIP6 models *Geophys. Res. Lett.* **47** e2019GL085782