Estimation of temperature and precipitation uncertainties using quantile neural networks

Andrew E. Brettin[1], **Laure Zanna**[1,2]

[1]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA
[2]Center for Data Science, New York University, New York, NY, USA

Key Points:

- We propose and evaluate a quantile neural network approach for constraining uncertainties on a variety of synthetic and observational datasets
- The quantile neural network's ease of implementation and generality reveal its suitability for quantifying uncertainties
- We compare the QNN against linear and Gaussian baselines, finding nonlinear dependencies for temperature and nonlinear and non-Gaussian dependencies for precipitation

Corresponding author: Andrew Brettin, `brettin@cims.nyu.edu`

Abstract

Given the challenges of limited predictability and risks that extreme events pose, imposing quantitative constraints on the variability of geophysical fields under observable but fluctuating conditions is necessary for assessing climate hazards. Here, we propose a quantile regression neural network framework for estimating uncertainties with two novel yet tractable modifications to the loss function to enforce uniform quantile accuracy and reduce the occurence of degenerate predicted probability distributions. We evaluate the quantile neural network against other conditional probabilistic baselines on a suite of datasets: synthetic datasets, observed in-situ daily temperature maxima from 1,501 NOAA Global Surface Summary of the Day (GSOD) weather stations, and altimetry-observed precipitation from Tropical Rainfall Measuring Mission (TRMM). On synthetic datasets, the proposed quantile regression neural network accurately predicts conditional distributions where more restrictive methods like linear quantile regression or mean-variance estimation neural networks are deficient, mitigates shortcomings of some other quantile neural network approaches, and converges stably under a range of different hyperparameters. Applying the quantile neural network to predict GSOD daily temperature maxima shows that temperature distributions may be relatively well-described as Gaussian, though nonlinear dependencies on the station sea level pressure and geopotential heights are likely important. On precipitation statistics, the quantile regression neural network strongly outperforms linear quantile regression and the Gaussian maximum likelihood network baselines, indicating it is able to capture the highly nonlinear and non-Gaussian conditional distributions of precipitation. The performance of the quantile regression neural network on a variety of datasets indicates that it is a simple, flexible, and general approach that could be used to constrain aleatoric uncertainties for a myriad of geophysical quantities which may have nonlinear or non-Gaussian conditional dependencies.

Plain Language Summary

The climate system is highly chaotic and unpredictable, often yielding extremes and risks that must be quantified. In light of these uncertainties, we propose a data-driven probabilistic technique, a type of "quantile neural network," for quantifying uncertainties which requires few assumptions and has a straightforard implementation. Using a synthetic dataset, we establish the advantages of this quantile neural network against baselines which require stronger assumptions, such as one that assumes linear relationships between inputs and outputs and one that assumes that uncertainties are normally-distributed. We then apply this technique to weather station temperature data and satellite observations of precipitation, finding that daily maximum temperatures are well-described by nonlinear relationships with normally-distributed uncertainties, whereas precipitation depends significantly nonlinearly on the inputs to the model and exhibits non-normal statistics. This work shows how quantile neural networks can be easily implemented to gain a more accurate representation of uncertainties in the geosciences.

1 Introduction

The climate system is governed by complex, highly nonlinear interactions between the atmosphere, ocean, land and cryosphere (Gupta et al., 2022), and the chaotic dynamics that result can be difficult to predict with certainty given limited information about the system state (Vitart et al., 2017). Interactions between processes occur over a range of temporal and spatial scales, producing variability and extremes, often with adverse impacts to human populations (Newman & Noy, 2023). Given the lack of predictability in the climate system, accurately quantifying the uncertainty of geophysical fields under changing measurable conditions is crucial due to the hazards that can result from its chaotic dynamics.

Often, uncertainties in the atmosphere-ocean system are represented using Gaussian statistics. Gaussian assumptions underlie climate projections (Kopp et al., 2014; IPCC, 2021), stochastic parameterization of sub-gridscale processes (Franzke et al., 2015), measurement approximations in data assimilation models and reanalysis products (Bocquet et al., 2010; Hersbach et al., 2020), linear inverse models (Penland, 1989; Penland & Sardeshmukh, 1995), in-situ and altimetric observational product quality control and more. The Gaussian distribution's relevance to characterizing a wide range of uncertainties in the Earth system can be attributed to a few theoretical considerations (Sura & Hannachi, 2015). Firstly, the normal distribution plays a crucial role in the Central Limit Theorem, which states that the sample mean of independently and identically distributed random variables approaches a normal distribution as the sample size increases. The Central Limit Theorem implies that aggregation operations (e.g., the averaging involved in the measurement, simulation, and forecasting of geophysical quantitites) tend to produce normally distributed quantities (DelSole & Tippett, 2022). Another reason that the Gaussian distribution arises is due to its unique role as the maximum entropy distribution for a given mean and variance for quantities with unbounded support (Sura & Hannachi, 2015; Majda & Wang, 2006). The principle of maximum entropy, first proposed by Jaynes (1957), is that the maximum entropy distribution (i.e. the least informative one under given constraints) is the most probable distribution. Thus, without further information constraining the form of the uncertainties, the Gaussian is the "best guess" of the underlying uncertainty.

One approach to quantifying Gaussian uncertainties has gained significant traction for geoscience applications in recent years is mean-variance estimation (MVE) neural networks. Proposed by Nix and Weigend (1994), MVE networks are data-driven models optimized over the Gaussian negative log-likelihood to yield not only a point estimate, but also a standard deviation to quantify the level of uncertainty given the inputs. MVE networks have been used for developing stochastic parameterizations (Guillaumin & Zanna, 2021; Perezhogin et al., 2023; Wu et al., 2025), identifying drivers of predictability (Gordon & Barnes, 2022), identifying exceedance times of critical global warming thresholds (Diffenbaugh & Barnes, 2023) and more (Haynes et al., 2023; Barnes & Barnes, 2021; Schreck et al., 2024).

Despite the proliferation of machine learning methods that assume Gaussianity, however, many geophysical quantities are non-Gaussian. Observations of surface air temperature exhibit significantly non-Gaussian characteristics (Fig. 1; Proistosescu et al. (2016); Catalano et al. (2021); Cavanaugh and Shen (2014); McKinnon et al. (2016)), with numerous physical causes (such as tracer advection-diffusion processes and jet dynamics) supported by numerical simulations (Linz et al., 2018; Garfinkel & Harnik, 2017; Hassanzadeh & Kuang, 2015) as well as theoretical arguments (Sura & Hannachi, 2015; Kimura & Kraichnan, 1993; McLaughlin & Majda, 1996; Hu & Pierrehumbert, 2002). Precipitation statistics are highly non-Gaussian, and modelling precipitation statistics remains an active area of research (Ashkenazy & Smith, 2024; Li et al., 2023; Scheuerer et al., 2020; Beck et al., 2020; Martinez-Villalobos & Neelin, 2019). Deviations from Gaussianity have implications for the quantification of extremes (Bjarke et al., 2023; Loikith & Neelin, 2019), as well as the changes in likelihood of tail events under changing temperatures (Loikith & Neelin, 2015).

Studies such as Barnes et al. (2023) have relaxed the Gaussian assumption of MVE networks by training neural networks to estimate parameters of probability distributions which include skewness and kurtosis parameters (e.g., the "SHASH" distribution; Jones and Pewsey (2019)). Such approaches generalize the types of aleatoric uncertainties that can be estimated; nevertheless, they require parametric assumptions about the underlying uncertainties which may not necessarily hold. For instance, the SHASH distribution may inadequately represent precipitation statistics, in particular due to the incompatability of nonnegative precipitation measurements with the SHASH's unbounded sup-

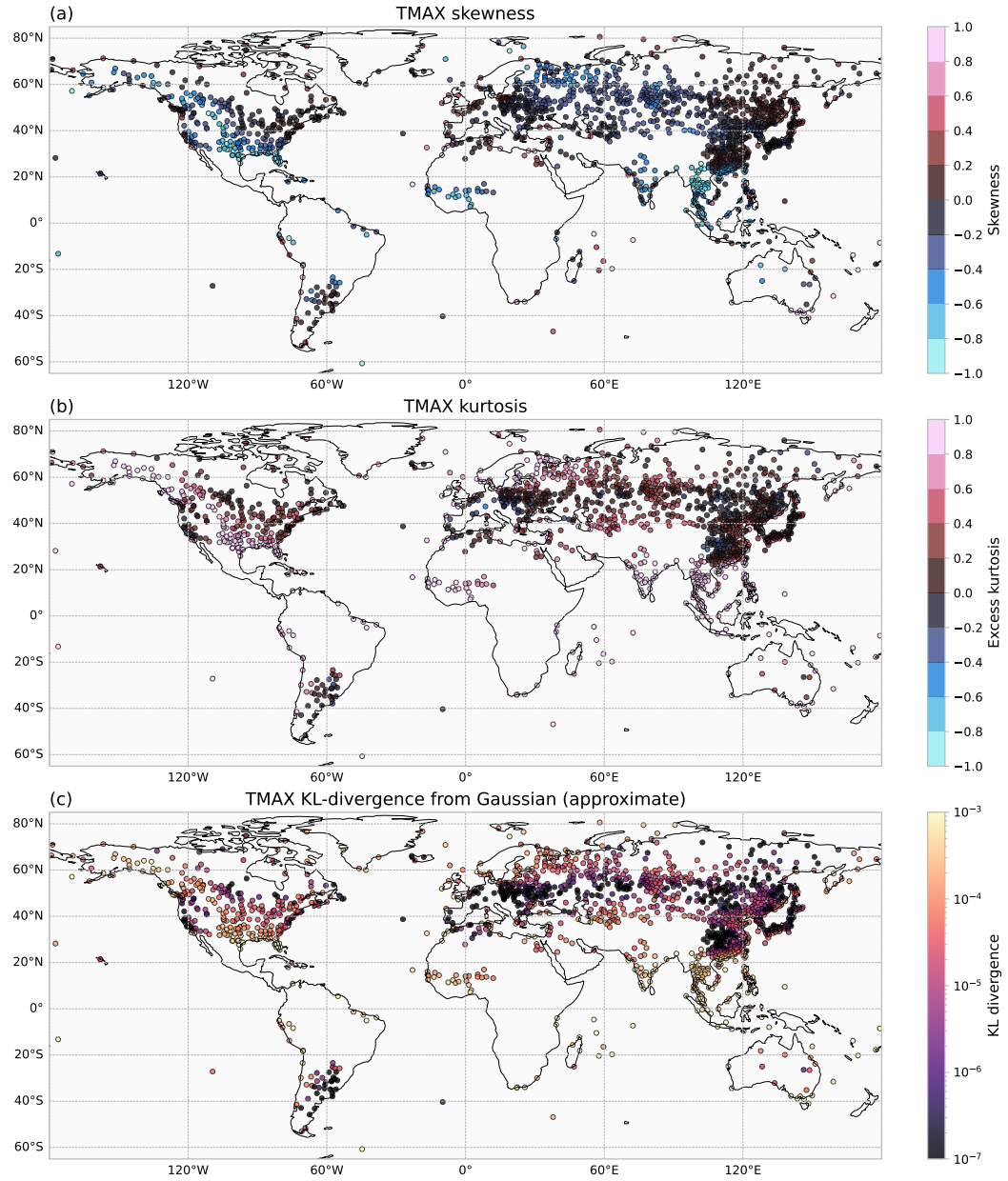

Figure 1. Non-Gaussianity of daily maximum surface air temperature (TMAX) seasonal anomalies, as measured by the sample's (a) skewness, (b) kurtosis, and (c) Kullback-Leibler divergence between the sample and a Gaussian estimated with the approach of Hyvärinen and Oja (2000) using contrast function $G(u) = \log\cosh u$ (higher values indicate greater deviations from Gaussianity).

port. Alternatively, quantile regression techniques offer ways to estimate the response distribution without parametric assumptions about the distribution of uncertainties by using the optimization formulation for the quantile (Koenker & Bassett Jr, 1978; Koenker, 2005). Linear quantile regression has been used for identifying temporal changes in distributions of surface air temperature observations (McKinnon et al., 2016) and for sea surface heights simulated by climate models (Falasca et al., 2023). The quantile regres-

sion loss has also been used to train regression neural networks, starting with White (1992) and Taylor (2000). Such quantile regression neural networks give an appealing means of estimating uncertainties, as they allow for complete estimation of nonlinear functional dependencies and non-Gaussian statistics. Several studies have used the quantile regression loss for neural networks to predict conditional probabilities, including for assessing financial risk (Chronopoulos et al., 2024; Keilbar & Wang, 2022), energy demands (W. Zhang et al., 2019; Tambwekar et al., 2021; Belloni et al., 2019), and healthcare outcomes risks (X. Zhang et al., 2025; Corsaro et al., 2024). Nevertheless, the use of quantile regression neural networks for geoscience applications seems to be limited to a few studies (Cannon, 2018; Haynes et al., 2023; Bremnes, 2020; Papacharalampous et al., 2025). Technical challenges, such as sample size limitations or complications associated with enforcing quantile monotonicity (Chernozhukov et al., 2010; Cannon, 2018; Padilla et al., 2022) may inhibit broader usage of quantile regression neural networks in the geosciences.

In this paper, we propose a flexible and simple quantile regression neural network for estimating uncertainties. Our implementation of the quantile regression neural network uses the Rectified Linear Unit (ReLU) as a loss function to encourage quantile monotonicity during training, and we refer to it as the "ReLU bias loss quantile neural network" (RBLQNN). We evaluate the RBLQNN against baselines which assume Gaussian conditional distributions or linear dependence to assess the relative importance of linearity or Gaussianity assumptions in a variety of synthetic and observational datasets. After establishing the advantages of the RBLQNN on synthetic datasets, we assess the Gaussianity and linearity assumptions of two observational datasets: NOAA Global Surface Summary of the Day (GSOD) daily temperature measurements at 1,501 weather stations, and Tropical Rainfall Measuring Mission (TRMM) precipitation altimeter observations.

In Section 2, we formulate our approach to conditional probability estimation, and describe the RBLQNN, baselines, datasets, and metrics. Then, in Section 3 we evaluate the performance of the RBLQNN, demonstrating its advantages over the baselines and other quantile regression neural network techniques. In Section 4, we examine the importance of Gaussianity and linearity assumptions in the GSOD and TRMM datasets. We end with a discussion providing some perspectives on our results and describing caveats of the RBLQNN in Section 5.

2 Methods

2.1 Conditional probability estimation

2.1.1 Formulation and optimization setup

In this framework, geophysical target variables Y such as surface air temperature or precipitation are considered a function r of random variables

$$Y = r(\mathbf{X}, \Psi), \tag{1}$$

where the $\mathbf{X} = (X_1, \ldots, X_p)$ represent observable random variates and Ψ represents the remaining aleatoric uncertainty. We seek to represent the distribution of the prediction variables conditioned on observed quantities $Y|\mathbf{X} = \mathbf{x}$.

We express probabilities in terms of quantiles. For a continuous cumulative distribution function (CDF) $F : \mathbb{R} \to (0, 1)$, the q-quantile y_q is defined by $y_q = F^{-1}(q)$. Because the CDF is unique for a given distribution, a probability distribution is fully characterized by the set of quantiles for $q \in (0, 1)$. Thus, the conditional distribution of $Y|\mathbf{X}$ can then be formulated in terms of its quantiles $(Y|\mathbf{X})_q$ by functions $f^{(q)} : \mathbb{R}^p \to \mathbb{R}$ as

$$y_q = (Y|\mathbf{X} = \mathbf{x})_q = f^{(q)}(\mathbf{x}). \tag{2}$$

Identifying the conditional distribution amounts to determining $f^{(q)}$ for $q \in (0, 1)$.

The key insight behind quantile regression techniques is that the q-quantile y_q satisfies the following optimization problem (Koenker & Bassett Jr, 1978):

$$y_q = F^{-1}(q) \iff y_q = \underset{u}{\operatorname{argmin}}\ \mathbb{E}[\rho_q(Y - u)] = \underset{u}{\operatorname{argmin}} \int_{\mathbb{R}} \rho_q(y - u)\ dF(y), \tag{3}$$

where the "pinball function" $\rho_q(t)$ is defined by

$$\rho_q(t) = qt\mathbb{1}_{\{t \geq 0\}} - (1 - q)t\mathbb{1}_{\{t<0\}}. \tag{4}$$

The pinball function is plotted for various values of q in Supporting Figure S1. Intuitively, the pinball function asymmetrically penalizes data above and below a given value to obtain estimates for a specified quantile. For instance, when predicting a high quantile like $q = 0.9$, data y is heavily penalized for exceeding the minimization argument u in (3), but lightly penalized for subordinating u. This pushes the optimal value of u towards the higher end of the data. When $q = 0.5$, the pinball function is symmetric and reduces to the absolute value of its argument (scaled by a factor of 0.5).

This optimization formulation (3) is analogous to how the mean of a distribution is the argument minimizer of the variance functional:

$$\mu = \mathbb{E}Y \iff \mu = \underset{u}{\operatorname{argmin}}\ \mathbb{E}[(Y - u)^2] = \underset{u}{\operatorname{argmin}} \int_{\mathbb{R}} (y - u)^2\ dF(y). \tag{5}$$

The optimization formulation (3) provides an avenue for determining conditional distributions using regression techniques. In a general regression problem, we seek to estimate some property of the conditional distribution of $Y|\mathbf{X}$ by a functional $f(\mathbf{X})$. This minimization formulation allows this functional to be empirically optimized over a class of candidate functions $\{f_\theta\}_{\theta \in \Theta}$ for a given dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ sampled from the joint distribution $(\mathbf{X}, Y)$. For example, in least-squares regression, the optimization formulation for the mean in (5) is used to empirically estimate the conditional mean:

$$\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = f(\mathbf{x}) \approx \hat{f}_\theta(x), \tag{6}$$

where

$$\hat{f}_\theta = \underset{f_\theta}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f_\theta(\mathbf{x}_i)\right)^2 \right\}. \tag{7}$$

In a similar vein, the optimization formulation for the quantiles (3) can be used to estimate the conditional quantile of the distribution:

$$y_q = (Y|\mathbf{X} = \mathbf{x})_q = f^{(q)}(\mathbf{x}) \approx \hat{f}_\theta^{(q)}(\mathbf{x}), \tag{8}$$

where

$$\hat{f}_\theta^{(q)} = \underset{f_\theta}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho_q \left(y_i - f_\theta(\mathbf{x}_i)\right) \right\}. \tag{9}$$

2.1.2 ReLU bias loss quantile regression neural network (RBLQNN)

The RBLQNN is a multilayer perceptron $f_\theta : \mathbb{R}^p \to \mathbb{R}^m$ that outputs predicted quantiles for a discrete set of probability levels $0 < q_1 < \cdots < q_m < 1$:

$$(\hat{y}_{q_1}, \ldots, \hat{y}_{q_m}) = f_\theta(\mathbf{x}). \tag{10}$$

Quantiles at intermediate probability levels $q \neq q_j$ can then be approximated using interpolation techniques. In this study, we evaluate predictions of the quantiles using $m =$

19 equispaced probability levels with increments of $\Delta q = 0.05$. This is a subjective choice loosely informed by statistical conventions (Fisher, 1970; Xu et al., 2017).

Since the neural network predicts multiple quantiles simultaneously, the loss function optimizes predictions for all quantiles:

$$\mathcal{L}_Q(y_i, \hat{\boldsymbol{y}}_{q,i}) = \sum_{j=1}^{m} \lambda_j \rho_{q_j}(y_i - \hat{y}_{q_j,i}). \tag{11}$$

We propose to combine two simple additions for mitigating issues that may arise when using this loss for training:

1. *Quantile counterbalancing.* At the quantile $y_q = F^{-1}(q)$, the expected value of the quantile loss $\mathbb{E}\left[\rho_q(Y - y_q)\right]$ is different for different probability levels q. In order to ensure that each quantile is optimized evenly, the loss weights λ_j in (11) should be chosen to be inversely proportional to the expectation $\lambda_j = \mathbb{E}\left[\rho_q(Y - y_q)\right]^{-1}$. This expectation reduces to
$$\mathbb{E}\left[\rho_q(Y - y_q)\right] = q\left(\mathbb{E}Y - \mathbb{E}[Y|Y < y_q]\right). \tag{12}$$
While the true expectation depends on the distribution of Y (which is unknown), Supporting Figure S2 shows that this expectation is not extremely sensitive to the type of distribution. For the purposes of developing bona fide weights, we assume a standard-normal distribution so that $\lambda_j = \exp\left(\frac{\Phi^{-1}(q_j)}{2}\right)$, where Φ is the CDF of the standard normal distribution.
2. *ReLU bias loss.* Because the CDF must be monotonic, probability distributions become degenerate when predicted quantiles cross. That is, for $q_1 < \cdots < q_m$, we must have that $\hat{y}_{q_1} < \cdots < \hat{y}_{q_m}$. Various methods for avoiding quantile crossings have been proposed, such as monotone rearrangement of predicted quantiles (Chernozhukov et al., 2010), applying monotonicity constraints to function inputs (Cannon, 2018) or predicting nonnegative increments for successive quantile values (Padilla et al., 2022). These methods range in complexity; here, we apply a simple bias loss to discourage quantile crossings during training while retaining the transparent output structure of the quantile regression neural network:
$$\mathcal{L}_{\mathrm{ReLU}}(\hat{\boldsymbol{y}}_{q,i}) = \sum_{j=1}^{m-1} \mathrm{ReLU}(\hat{y}_{q_j,i} - \hat{y}_{q_{j+1},i}), \tag{13}$$
where the Rectified Linear Unit function is given by $\mathrm{ReLU}(x) = x\mathbf{1}_{\{x \geq 0\}}$. If any of the quantiles is greater than the succeeding quantile, the ReLU loss is positive. If all quantiles are ordered, then the ReLU penalty is 0.

The net loss function that is used for training the RBLQNN is given by

$$\mathcal{L}(y_i, \hat{\boldsymbol{y}}_{q,i}) = \mathcal{L}_Q(y_i, \hat{\boldsymbol{y}}_{q,i}) + \eta \mathcal{L}_{\mathrm{ReLU}}(\hat{\boldsymbol{y}}_{q,i}), \tag{14}$$

where $\eta > 0$ is a hyperparameter.

2.1.3 Baselines

The predicted distributions made by the quantile neural network are compared against two baselines: linear quantile regression (LQR) and mean-variance estimation (MVE) neural networks. Linear quantile regression allows for arbitrary conditional probability estimates, but the functional dependence on the regressors is assumed to be linear. On the other hand, mean-variance estimation networks allow for nonlinear dependencies, but restrict conditional probabilities to be Gaussian. Together, these baselines help assess the relative importance of nonlinearities and non-Gaussianity in the predicted conditional distributions.

2.1.3.1 Linear quantile regression (LQR) For linear quantile regression, it is assumed that the quantiles of the regressand depend linearly on the regressors:

$$y_q = (Y|\mathbf{X} = \mathbf{x})_q = \beta_0(q) + \beta(q)^T\mathbf{x}. \tag{15}$$

Accordingly, the family of functions minimized using (7) are limited to linear functions:

$$\hat{\beta}_0(q), \hat{\beta}_1(q) = \underset{\beta_0(q),\beta_1(q)}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \rho_q(y_i - \hat{y}_{q,i}), \tag{16}$$

where $\hat{y}_{q,i} = \beta_0(q) + \beta_1(q)x_i$.

For the linear quantile regression model, predictions are made for each quantile $q_1, \ldots, q_m$ simultaneously, so that weights are shared for different quantile predictions, which has been shown to improve predictions (Zou & Yuan, 2008; Jiang et al., 2012).

2.1.3.2 Mean-variance estimation (MVE) neural networks As an alternative to linear quantile regression, we explore predictions made by mean-variance estimation (MVE) networks (Nix & Weigend, 1994). MVE networks are artificial neural networks that yield predictions of conditional probability densities by outputting the parameters of a normal distribution $\hat{\mu}_i = \hat{\mu}(\boldsymbol{x}_i), \hat{\sigma}_i^2 = \hat{\sigma}^2(\boldsymbol{x}_i)$. (In practice, because the predicted standard deviation should be positive, the log-variance $\log(\hat{\sigma}_i^2)$ is typically used as an output of the neural network and then exponentiated to yield positive values of $\hat{\sigma}_i^2$.) Weights and biases in the neural network are updated by optimizing the log-likelihood of the target data under the predicted parameters:

$$\mathcal{L}\left(y_i, \hat{\mu}(\boldsymbol{x}_i), \hat{\sigma}^2(\boldsymbol{x}_i)\right) = -\log p(y_i|\hat{\mu}_i, \hat{\sigma}_i^2) = \frac{1}{2}\left[\log(\hat{\sigma}_i^2) + \left(\frac{y_i - \hat{\mu}_i}{\hat{\sigma}_i}\right)^2\right] + C, \tag{17}$$

where $p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$ is the probability density function of a normal distribution with parameters μ and σ^2, and $C = \frac{1}{2}\log 2\pi$ is an immaterial optimization constant. Quantile predictions are then given by $\hat{y}_q = \hat{\mu}_i + \hat{\sigma}_i\Phi^{-1}(q)$.

2.1.3.3 Alternative quantile regression neural network approaches We also assess the performance of our quantile regression neural network trained on (14) against various other quantile regression neural network techniques:

1. *Unweighted network.* This quantile neural network is trained using (11) with uniform weighting $\lambda_1 = \cdots = \lambda_m = 1$, and is termed the "composite" network in Xu et al. (2017).
2. *No-bias network.* This network uses the normal-distribution inverse-expectation weighting scheme $\lambda_j = \exp\left(\frac{\Phi^{-1}(q_j)}{2}\right)$. However, the bias loss of (13) is not applied.
3. *Cumulative increment.* Proposed in Padilla et al. (2022), this network enforces strict monotonicity by predicting positive increments between successive quantiles. Specifically, the network outputs values $h_1(\mathbf{x}), \ldots, h_m(\mathbf{x})$ such that

$$\hat{y}_{q_j} = \begin{cases} h_1(\mathbf{x}) & (j = 1), \\ h_1(\mathbf{x}) + \sum_{k=2}^{j}\log(1 + e^{h_k(\mathbf{x})}) & (j = 2, \ldots, m). \end{cases} \tag{18}$$

   Since $\log(1 + e^t) > 0$, monotonicity of predicted quantiles is strictly enforced. This network can then be trained using the loss function in (11) using equal weights $\lambda_1 = \cdots = \lambda_m = 1$.

2.1.3.4 Mean squared error (MSE) neural networks In addition to the probabilistic baselines, a neural network trained using the Mean Squared Error (MSE) to predict the target variable is also used as a deterministic baseline.

2.2 Datasets

We evaluate the methods for conditional probability estimation on a suite of different datasets: 1) synthetic datasets, 2) weather station daily temperature maxima, and 3) altimetry-observed precipitation.

2.2.1 Synthetic datasets

We first demonstrate the performance of the different methods on three different synthetic datasets:

- *Dataset 1.* $Y = X^2 + \Psi$, where $X \sim N(0,1)$ is drawn from a standard normal and Ψ is drawn from a Gumbel distribution. The Gumbel distribution, which is used in extreme value theory, was selected to represent the aleatoric uncertainty due to its non-Gaussianity and support on the entire real line.
- *Dataset 2.* $Y = \text{Beta}(\alpha, \beta)$, where $\alpha = X + 0.2$, $\beta = 1.2 - X$, and $X \sim U(0,1)$ is uniform distributed on $[0,1]$. This dataset was formulated to investigate the impacts of heteroscedasticity on the conditional probability estimation techniques. The Beta distribution has bounded support, providing an interesting test case for the RBLQNN and other baselines.
- *Dataset 3.* Often, uncertainties of interest arise from quantities generated as a result of a time-varying dynamic process. Thus, we consider a stationary distribution generated from a two-dimensional stochastic gradient system of damped Langevin dynamics (Schlick, 2010) of $\boldsymbol{x} = (x, y)$,

$$\frac{d\boldsymbol{x}}{dt} = -\nabla V(\boldsymbol{x}) + \xi, \tag{19}$$

  where $V(\boldsymbol{x})$ is a potential function and $\xi \sim N(0,1)$ is white noise. The stationary joint distribution f yielded by the dynamics is governed by the steady-state Fokker-Planck equation, with solution given by the Boltzmann distribution $f(x,y) \propto e^{-V(x,y)}$ (Landau & Lifshitz, 2013). The conditional distributions can then be computed using numerical integration. We consider a potential of the form

$$V(x,y) = \prod_{i=1}^{3} \left((x - x_i)^2 + (y - y_i)^2 \right) \tag{20}$$

  which contains local minima at the points (x_i, y_i). For this dataset, we set $(x_1, y_1) = (-\frac{1}{2}, -\frac{1}{2})$, $(x_2, y_2) = (1, -1)$, and $(x_3, y_3) = (1, \frac{1}{2})$, which yields a unimodal conditional distribution at $x = x_1$ and a bimodal conditional distribution at $x = x_2 = x_3$. We simulate the trajectories from the origin using an Euler-Maruyama solver with timesteps $\Delta t = 10^{-4}$ up to a final time of $T = 50,000$. The joint distribution and corresponding histogram with 10,000 bins (100 equispaced bins in x and y) are shown in Supporting Figure S3. Histogram frequencies are within 2% of the true density for all bins, indicating that the samples are consistent with the theoretical distribution yielded by the Fokker-Planck equation. In order to verify that samples are drawn from the stationary distribution, Supporting Figure S4 shows the ensemble mean and spread of the magnitude $\|(x,y)\|_2$ to show that initial condition information is lost after a few seconds of simulation time. The training, validation, and testing sets are randomly sampled from the values generated by this timeseries.

Figures 2a, 2e, and 2i show the quantiles of the conditional distribution of $Y|\mathbf{X}$ and samples drawn from the joint distribution $(\mathbf{X}, Y)$ for the three synthetic datasets. For each dataset, we examine performance using training sets of $n = 10,000$ samples, validating training on a dataset of $1,000$ samples and measuring performance on a test dataset of $1,000$ samples. The input and target samples of Datasets 1 and 3 are standardized

using the training dataset mean and standard deviation. No standardization or normalization was applied to Dataset 2, as the samples already take values between 0 and 1.

2.2.2 NOAA Global Surface Summary of the Day daily temperature maxima

We next consider temperature observations from 1,501 weather stations over 1960–2020 given by the National Oceanic and Atmospheric Administration's (NOAA) Global Surface Summary of the Day (GSOD) dataset (NOAA National Centers of Environmental Information, 1999). The GSOD dataset contains daily summary statistics for 18 surface meteorological variables derived from the NCEI Integrated Surface Hourly (ISH) dataset (Lott et al., 2001; A. Smith et al., 2011). We use the daily maximum temperatures (TMAX) and daily-mean sea level pressures (SLP) from this dataset. In addition, we supplement the GSOD dataset with 500 mb and 850 mb geopotential heights (z500, z850) from ECMWF's ERA5 reanalysis (Hersbach et al., 2020). These variables were included to provide synoptic information about the atmospheric circulation at the middle and lower troposphere. For each weather station, models are fitted to predict TMAX conditioned on concurrent station-derived SLP and geopotential heights from the nearest ERA5 gridpoint. Separate models are trained for each station.

The GSOD dataset contains records from more than 9,000 stations. We filter data-limited stations by requiring a minimum number of observations for training, validation, and testing. We use data from years 1960–2010 for training, 2011–2015 for validation, and 2016–2020 for testing, and stations are rejected if there are fewer than 30 years of daily records for training (10,950 samples) or 3 years for validation or testing (1,095 samples). This leaves 1,501 stations for our analysis. Inputs and targets are not detrended or deseasonalized, but values are standardized according to the mean and standard deviation of the training data for each station.

2.2.3 TRMM altimetry precipitation observations with ERA5 reanalysis

As an additional test case, we use reanalysis and altimetry observations of atmospheric predictors to predict altimetry-observed precipitation levels. The samples in the dataset consist of hourly snapshots in a $2^\circ \times 2^\circ$ box along swaths measured by the Tropical Rainfall Measuring Mission (TRMM) satellites from 2000–2010 (Kummerow et al., 2000). As predictors from ERA5, we use equivalent potential temperatures averaged over the planetary boundary layer (1000–900mb) and free troposphere (850–400mb), as well as column relative humidities and precipitable water integrated over the planetary boundary layer and free troposphere. Additionally, we use convective available potential energy, surface air temperatures, 500-hPa vertical velocities, and entire-troposphere precipitable water from ERA5. Finally, the subgrid percentage of convective and stratiform areas measured by TRMM radar are used, as they are a leading indicator of precipitation amount (Ahmed & Schumacher, 2015). The first 80% of observations are used for training (176,224 samples), while the validation and testing set consist of the penultimate and final 10% of samples, respectively (22,028 samples).

2.3 Metrics

We employ several metrics to assess our predictions:

1. *Discrimination metrics.* The Mean Absolute Error (MAE) between the true quantiles at probability level q ($\mathbf{y}_q$) and prediction ($\hat{\mathbf{y}}_q$) is computed by

$$\mathrm{MAE}(\mathbf{y}_q, \hat{\mathbf{y}}_q) = \frac{1}{n} \sum_{i=1}^{n} |y_{q,i} - \hat{y}_{q,i}|. \tag{21}$$

Low quantile prediction MAE implies that forecasts are correctly approximating the conditional probability distribution. However, it requires that the ground truth distribution is available.

2. *Calibration metrics.* Calibration, also sometimes referred to as reliability, measures the probability of forecasted events against the frequency of those events (Dawid, 1982; Gneiting et al., 2007). To measure calibration, the Probability Integral Transform (PIT) (Dawid, 1984) is useful. The premise of the PIT is that the predicted cumulative distribution $\hat{F}_i$ evaluated at the observed value y_i should follow a uniform distribution. Thus, the histogram of $p_i = \hat{F}_i(y_i)$ should be approximately flat, and the deviations from a flat histogram can be measured against the expected level of deviation for a truly uniform distribution. The prediction of different quantiles also lends itself naturally to computing PIT histograms. Consider m uniform quantile predictions $\hat{y}_{q_1,i}, \ldots, \hat{y}_{q_m,i}$ for a given sample i. This partitions the real line into $B = m + 1$ bins $B_k = [\hat{y}_{q_k,i}, \hat{y}_{q_{k+1},i})$ for $k = 0, \ldots, m$, using the convention that $\hat{y}_{q_0,i} = -\infty$ and $\hat{y}_{q_{m+1},i} = \infty$. Given samples $i = 1, \ldots, n$, let n_k be the total number of observations y_i that fall in B_k. The deviation statistic is given by

$$D = \sqrt{\frac{1}{B}\sum_{k=0}^{m}\left(r_k - \frac{1}{B}\right)^2}, \tag{22}$$

where $r_k = \frac{n_k}{n}$ gives the frequency of the k^{th} bin. This deviation statistic is measured against the expected level of deviation for a true uniform distribution (Bourdin et al., 2014)

$$\mathbb{E}D = \sqrt{\frac{1 - B^{-1}}{nB}}. \tag{23}$$

Alternatively, the likelihood of the deviation statistic D can also be quantified under the null hypothesis that the PIT histogram is sampled from a uniform distribution. Under the limit of large sample size n and bin counts n_k, the deviation statistic for samples from a uniform distribution has a chi-squared distribution (Wasserman, 2004):

$$nB^3D^2 \sim \chi^2(B - 1). \tag{24}$$

Thus, the deviation statistic can be used to test the null hypothesis of uniformity against the alternative that the PIT histogram is not uniform (and that the predicted distributions are not well-calibrated).

3. *Proper scores.* We also consider the continuous ranked probability score (CRPS; Matheson and Winkler (1976)). The CRPS measures the quality of a probabilistic prediction $\hat{F}$ against an observed value y. The CRPS is given by

$$\text{CRPS}(\hat{F}, y) = \int_{-\infty}^{\infty}\left(\hat{F}(x) - \mathbb{1}_{\{x>y\}}\right)^2 dx. \tag{25}$$

The CRPS is a strictly proper scoring rule, meaning that on expectation, it is optimal when $\hat{F}$ is the true sampling distribution for the observations y (Gneiting & Raftery, 2007). Thus, the CRPS cannot be improved by hedging for alternative outcomes. Predictions are penalized for being overconfident as well as underconfident; as such, the CRPS assesses both the calibration and the sharpness of the predicted distributions simultaneously. Notably, the CRPS is a generalization of the Mean Absolute Error in the case in which predictions are point estimates and $\hat{F}(x)$ is a Heaviside function. The CRPS can be approximated directly from the predicted quantiles by using the quantile formulation of the CRPS from Laio and Tamea (2007) and employing numerical approximations (Bröcker, 2012; Taggart, 2023). Under equispaced quantile predictions, this is evaluated as

$$\text{CRPS}(\hat{\mathbf{y}}_q, y) = \frac{2}{m}\sum_{j=1}^{m} q(y - y_{q_j})\mathbb{1}_{\{y>y_{q_j}\}} + (1-q)(y_{q_j} - y)\mathbb{1}_{\{y\leq y_{q_j}\}}, \tag{26}$$

which is simply double the average pinball loss of the observation under the given quantiles.

2.4 Optimization and training procedure

For an appropriate comparison between the RBLQNN and other baselines, similar architectures and hyperparameter configurations are used for all models for each dataset considered in this study. The hyperparameter configurations for each dataset are given in Supporting Table S1. A hyperparameter sweep over different learning rates, model sizes, and regularizations for the synthetic datasets (Fig. 5) indicated that models converged well under similar configurations for the RBLQNN and baselines, justifying the use of identical learning rates and network sizes for comparison of different model types. Model weights are optimized using the Adam optimizer (Kingma & Ba, 2014). Weights are saved as checkpoints at the end of each epoch during training, and the checkpoint with the best loss over the validation set is used for analysis on the test set. Early stopping is also employed to limit computational costs from training many thousands of networks.

MVE networks are known to have some stability issues due to gradients typically being more sensitive to the predicted standard deviation than the predicted mean (Sluijterman et al., 2024; Seitzer et al., 2022). Therefore one recommendation from Nix and Weigend (1994) is to employ a "warm-up period" where only the mean is optimized but the variance is fixed over the first few epochs. To enforce training stability, during the initial warmup phase a fixed variance σ_F^2 is prescribed in equation (17), and outputted predicted log-variances are penalized for deviations from this fixed variance using an MSE loss:

$$\mathcal{L}\left(y_i, \hat{\mu}(\boldsymbol{x}_i), \hat{\sigma}^2(\boldsymbol{x}_i)\right) = \frac{1}{2}\left[\left(\frac{y_i - \hat{\mu}_i}{\sigma_F}\right)^2\right] + (\log\hat{\sigma}_i^2 - \log\sigma_F^2)^2, \tag{27}$$

where the first term is taken from the Gaussian log-likelihood loss (17) and the second term penalizes predicted log-variances, respectively.

The RBLQNN was found to converge more stably than the MVE networks without any such warm-up period. However, including an analogous warm-up phase for the RBLQNN tended to result in sharper conditional probability estimates. Thus, for the RBLQNN, during the same initial warm-up epochs used for the MVE networks, an MSE loss is used between the outputs of the RBLQNN and training targets. After the warm-up period is complete, the weights are adjusted using the loss (14) to predict the quantiles of the distribution.

A summary of the hyperparameter configurations for each dataset is given in Supporting Table S1.

3 Model performance on synthetic data

3.1 Model error

Figure 2 shows the predicted quantiles of the different conditional probability estimation techniques on the three synthetic datasets described in Section 2.2.1. Signed errors $\hat{y}_q - y_q$ in the quantile predictions as a function of x are given in Supporting Figure S5. These datasets clearly illustrate the limitations of LQR and the MVE neural networks: LQR fails to estimate highly nonlinear functional dependencies (e.g. Figure 2b), while MVE networks fail to capture non-Gaussian conditional distributions (Fig. 2g). However, even in situations where the assumptions of linearity or Gaussianity offer decent but imperfect approximations, the RBLQNN results in better estimates of the conditional distributions. For instance, although the normal distribution and Gumbel distribution share certain features such as unimodality and infinite support (Fig. 2a), the

RBLQNN is able to capture the skewness of the distribution (Fig. 2d) while the MVE neural network cannot (Fig. 2c). Similarly, although the linear quantile regression accurately approximates the nearly linear median quantiles of Dataset 2 (Fig. 2e and 2g), the quantiles at the tails of the distribution are poorly predicted by the linear quantile regression model while they are well-fit for the RBLQNN (Fig. 2h).

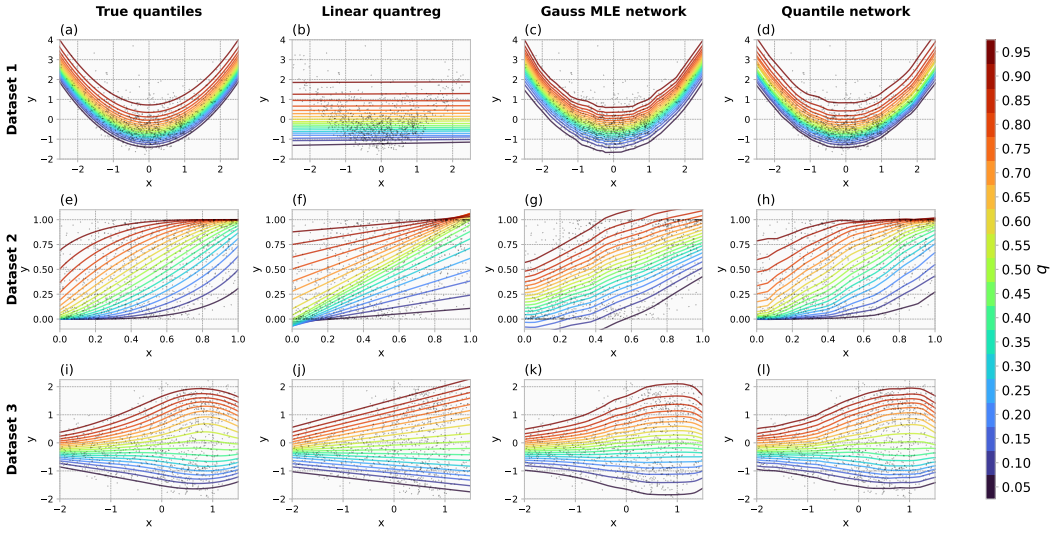

Figure 2. Predicted conditional quantiles on Synthetic Dataset 1 (a–d), Dataset 2 (e–h), and Dataset 3 (i–l). Scatterplot shows the test sample generated from the ground truth distribution, while colored lines indicate predicted quantiles at specified probability levels q. (a, e, i) True quantiles. (b, f, j) Predictions made using linear quantile regression. (c, g, k) Predictions made using MVE networks. (d, h, l) Predictions made using the RBLQNN.

Figure 3 shows the average errors of each of the individual quantile predictions averaged over the test set and further demonstrates how the assumptions made by each baseline result in poorly predicted distributions. Dataset 1 shows that inadequately capturing nonlinear dependencies with linear quantile regression can result in systematically large errors for all quantiles in the predicted conditional distribution (Fig. 3a). The MVE network resolves the issue of nonlinearity, but the assumption of Gaussianity can result in poor estimation of particular quantiles. For instance, while the MVE network provides reasonable estimates for the bulk of the distribution in Dataset 1, the tails of the distribution tend to be underestimated (Fig. 2c and Fig. 3a). Large errors in the tails of the distribution also occur in Dataset 2 and 3, as the Gaussian approximation is unable to capture the compact support of the distribution for Dataset 2 nor the bimodality of Dataset 3. The bimodality of Dataset 3 also results in the MVE network producing poor quantile predictions near each lobe of the bimodal conditional distribution (Fig. 2k and 3c). In contrast with each of these baselines, the RBLQNN tends to produce quantile predictions with good predictions for all quantiles of the distribution.

3.2 Comparison against other quantile regression neural network approaches

Figure 4 compares the performance of the RBLQNN to the alternative quantile regression neural network approaches given in Section 2.1.3. Since the network performance is sensitive to the weight initialization, for each quantile neural network method, an en-

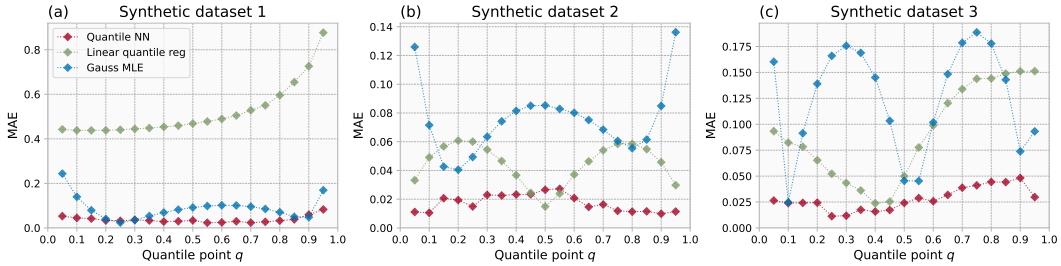

Figure 3. Mean absolute errors of predicted quantiles for the RBLQNN (red), linear quantile regression (light green), and MVE neural network (blue) for (a) Synthetic Dataset 1 and (b) Synthetic Dataset 2.

semble of 100 networks with weights initialized using different random seeds is used to robustly evaluate performance.

For the three synthetic datasets, the RBLQNN tends to produce better quantile predictions than the baseline quantile neural network techniques (Fig. 4a, 4c, and 4e). The RBLQNN error is similar to, though on average slightly lower than, the quantile neural networks which directly predict quantiles (the "unweighted" and "no-bias" networks). This seems to be primarily due to the inclusion of the ReLU bias loss (Eq. 13), as the comparisons between the two baselines with different quantile weighting schemes do not result in substantially different performance. Including the ReLU bias loss to encourage predicted distributions to be valid may impose stability constraints which facilitate network convergence during training.

Figure 4b, 4d, and 4f show the proportion of test samples which result in monotonic quantile predictions (i.e., nondegenerate conditional probabilities). The RBLQNN results in significant reductions in the number of predicted conditional distributions that are degenerate when compared to the unweighted and no-bias networks. For instance, applying the ReLU bias loss in Dataset 2 reduces the average number of samples with quantile crossings from more than 25% to less than 3% (Supporting Table S2). For Dataset 3, all but two ensemble members completely eliminate quantile crossings with the RBLQNN, whereas roughly 20% of ensemble members have quantile crossings for the unweighted and no-bias quantile networks (Supporting Table S3).

While the RBLQNN reduces the number of quantile crossings compared to other quantile neural networks which directly predict quantiles, it does not completely eliminate them. In contrast, the cumulative increment network's design explicitly prohibits all quantile crossings. However, Figures 4a, 4c, and 4e all show that the predicted distributions made using the cumulative increment network tend to be worse than the quantile neural networks which directly output quantiles. The relatively poorer performance of the cumulative increment network is likely due to instabilities resulting from the sum in (18) used for predicting higher order quantiles.

3.3 Training stability

A central challenge in the implementation of neural networks is that model performance can be sensitive to various hyperparameters (Goodfellow et al., 2016). Weights and biases in a neural network are normally optimized using stochastic gradient-based optimization techniques, which can evolve unpredictably under the highly nonconvex, high-dimensional loss landscapes set by the training data, network architecture, regularization, and loss functional. The convergence of a neural network thus depends on all of the factors that determine the loss landscape as well as the specifications that pre-

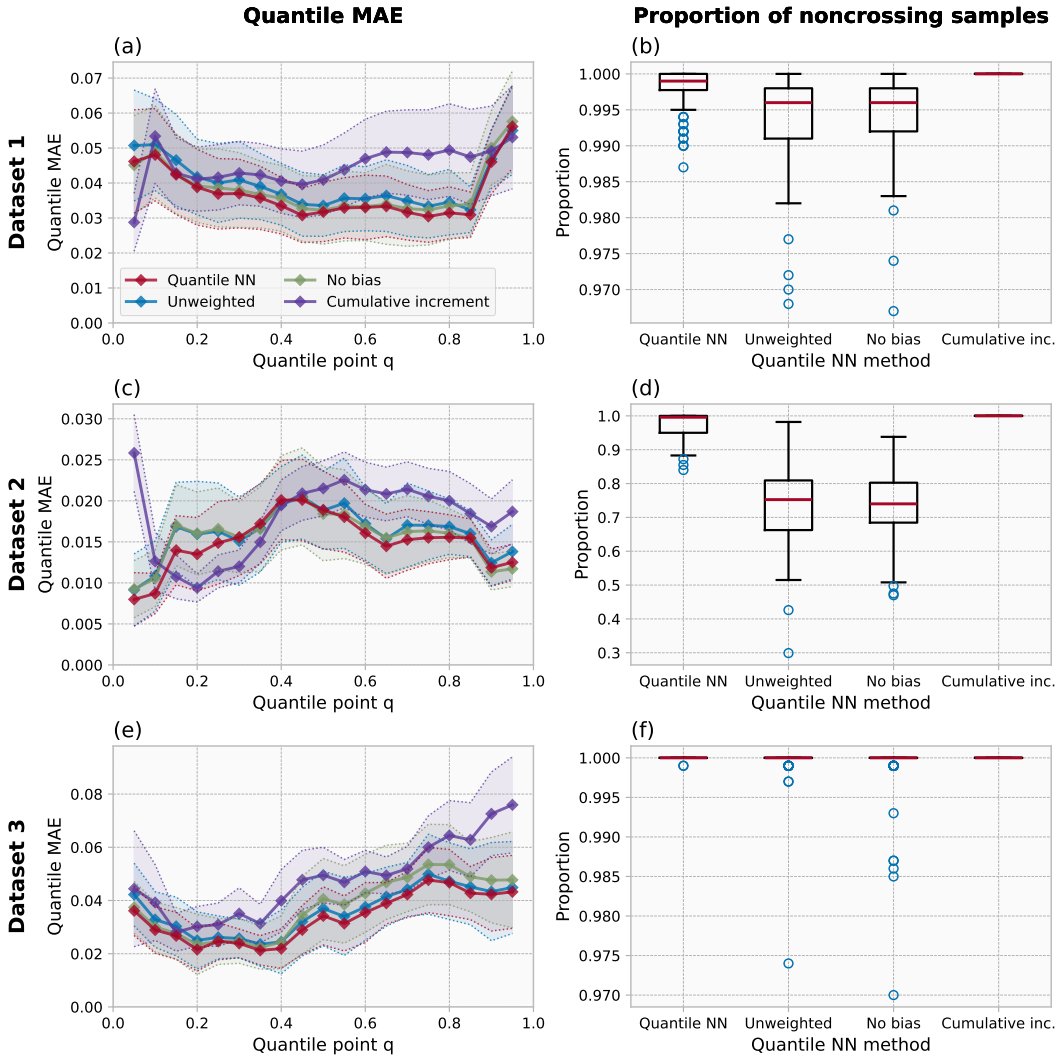

Figure 4. Performance of different quantile regression neural network techniques on Synthetic Dataset 1 (a, b), Dataset 2 (c, d), and Dataset 3 (e, f). (a, c, e): Mean absolute errors for quantile neural network predictions, averaged over 100 ensembles created by initializing network weights with different random seeds. Red: RBLQNN (Eq. 14). Blue: Unweighted quantile regression neural network of Xu et al. (2017). Light green: No-bias quantile regression neural network framework implementing the inverse-expectation weighting but without the ReLU bias loss (Eq. 13). Purple: cumulative increment network of Padilla et al. (2022). (b, d, f) Boxplots showing the fraction of nondegenerate probability distributions predicted by each quantile neural network approach. Conditional probability distributions are considered degenerate for a given sample if the predicted quantiles are not monotonic.

scribe the optimization procedure. As a result, hyperparameter selection often requires extensive tuning or automated sweeps over various configurations to obtain a well-fitted model. It is therefore advantageous to train using a loss function that tends to converge robustly over a range of different hyperparameters and datasets. We investigate the sensitivity of the RBLQNN against the MVE networks for the three synthetic datasets.

Network training is often particularly sensitive to hyperparameters such as the learning rate, network architecture, and regularization (L. N. Smith, 2018; Godbole et al., 2023). Therefore, we apply a grid search of different hyperparameters over a range of learning rates, regularizations, and network sizes with values provided in Supporting Table S4. We use eight different learning rates, three regularization levels, 2 layer sizes and 3 different numbers of neurons per layer, to sample a total of 144 hyperparameter combinations. All other hyperparameters are set using the configuration given in Supporting Table S1.

Figure 5a shows the best validation loss attained during training by the RBLQNN and MVE network for each of the 144 hyperparameter combinations. To permit a comparison between the different loss functions for the different networks, validation losses are normalized to the $[0, 1]$ range so that the hyperparameter configuration with the lowest loss over the 144 hyperparameter combinations has a normalized value of 0 and the highest loss has a normalized value of 1. Kernel density estimates indicate that losses cluster near the minimum values, indicating convergence for a broad range of hyperparameters for both the RBLQNN and MVE networks. However, for each of the three datasets, more values cluster near the minimum loss for the RBLQNN than for the MVE network, suggesting that training tends to be more stable for the RBLQNN than for the MVE network. Notably, the RBLQNN appears to attain strong performance for a broader range of learning rates, a crucial hyperparameter for neural network training. For low learning rates (e.g. 10^{-8}), normalized validation losses are high for each network; however, increasing learning rates results in better performance gains for the RBLQNN than for the MVE network (e.g., a normalized validation loss that is 18.2% lower on average for the RBLQNN than MVE network when using learning rates of 10^{-7}). For high learning rates (10^{-1}), losses begin to increase significantly for the MVE network but less so for the RBLQNN.

Figure 5b, 5c, and 5d show the empirical cumulative distribution function of the losses given in the strip plots in Figure 5a for each of the three synthetic datasets. The CDF of the RBLQNN is above the CDF of the MVE network at most loss levels, indicating that a greater proportion of RBLQNN converge within a given margin of the lowest loss than the MVE networks. For instance, for Synthetic Dataset 1, only 66% (71%) of MVE neural networks converge within 5% (20%) of the minimum loss, whereas 69% (78%) of the RBLQNNs converge within this margin of the minimum loss.

3.4 Sample-based metrics

The low MAE of the predicted quantiles made on the synthetic datasets indicates that the RBLQNN can successfully approximate conditional distributions. However, for most datasets, the ground truth distribution is unknown, making it impossible to assess predicted conditional probabilities directly against the true distribution. In this section, we explore sample metrics that assess the predicted distributions against the data.

Figures 6a, 6c, and 6e show histograms of the CRPS yielded by the predicted distributions of the different conditional probability estimation techniques over the test set. Averaged over all test samples, the RBLQNN has the lowest average CRPS of the three conditional probability estimation techniques for all synthetic datasets. The sample average CRPS for the RBLQNN nearly matches the CRPS obtained by using the ground truth conditional probability distribution for all three datasets.

Despite the RBLQNN attaining a lower sample-averaged CRPS for all three datasets, the histograms of CRPS significantly overlap for each of the different methods. Since the CRPS is a sample-based metric which is optimal only on expectation, the statistical significance of the low sample-averaged CRPS must be assessed. To assess the statistical significance of the sample-averaged CRPS over a range of different sample sizes, we employ bootstrapping with 100 ensemble members for a variety of bootstrap sample sizes

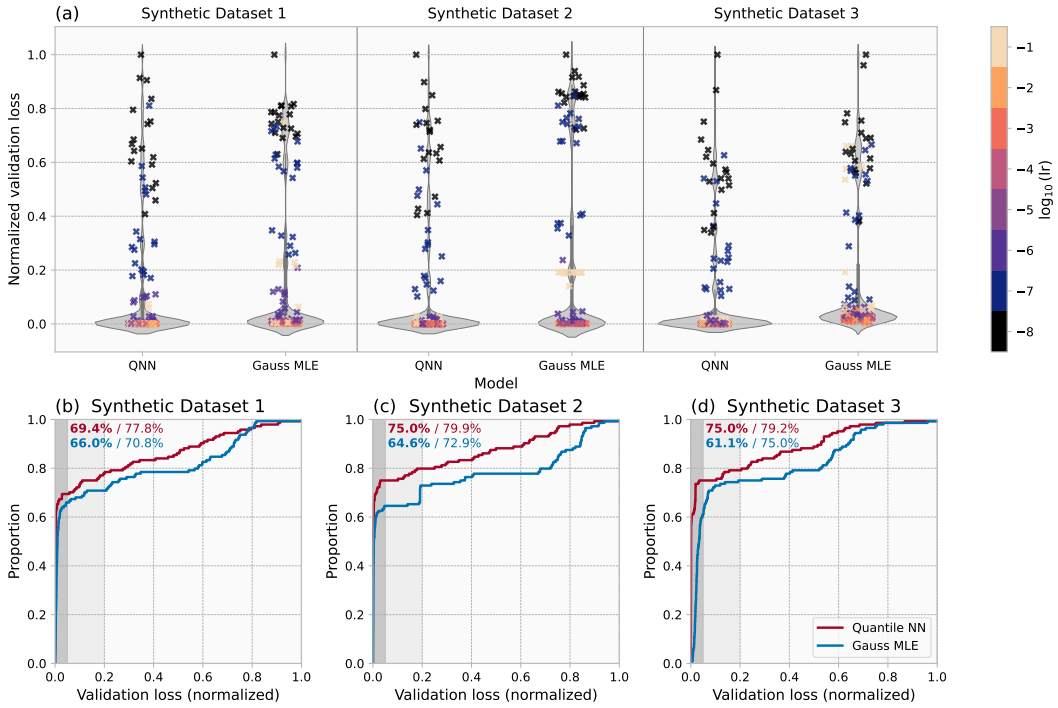

Figure 5. Training stability of the RBLQNN and MVE network evaluated over 144 hyperparameter combinations for the three synthetic datasets. (a) Strip plots of the lowest validation loss attained during training for the RBLQNN and MVE for the three different datasets, with gray shading showing a kernel density estimate. Losses are normalized to the $[0, 1]$ range. Colors indicate the learning rate. (b, c, d) Empirical CDFs of the normalized validation losses for each of the three different synthetic datasets for the RBLQNN (red) and MVE network (blue). Boldface percentages in the top left corner indicate proportion of neural networks which converge within 5% of the best loss from all hyperparameter configurations (indicated by dark gray shading). Plain typeface percentages indicate proportions of neural networks within 20% of the best loss (light gray shading).

N_b from 10 to 10,000. For each ensemble member, N_b samples are drawn with replacement from the test set, and the sample-averaged CRPS is computed over that sample for the various conditional probability estimation techniques. The CRPS sample averages for different ensemble members are compared pairwise with the true distribution CRPS sample averages to evaluate whether the predictions made by a given method are statistically distinguishable from the true distribution for a given sample size N_b. For instance, if among the 10,000 pairwise comparisons between the 100 ensemble members of the MVE network and 100 members of the true distribution CRPS sample averages fewer than 5% of the comparisons yield lower CRPS sample means for the MVE network, then it can be concluded that the probabilistic predictions made by the MVE network are inadequate for that sample size.

Figures 6b, 6d, and 6f show the ensemble spread of the CRPS sample averages for each of the different methods for a variety of bootstrap sample sizes N_b. As the bootstrap sample size increases, the ensemble spread of the CRPS sample averages decreases for each of the different methods, revealing the minimum sample sizes needed to reject the probabilistic models. For instance, for Dataset 1 only a few hundred samples are needed to establish that the linear quantile regression poorly predicts probability distributions,

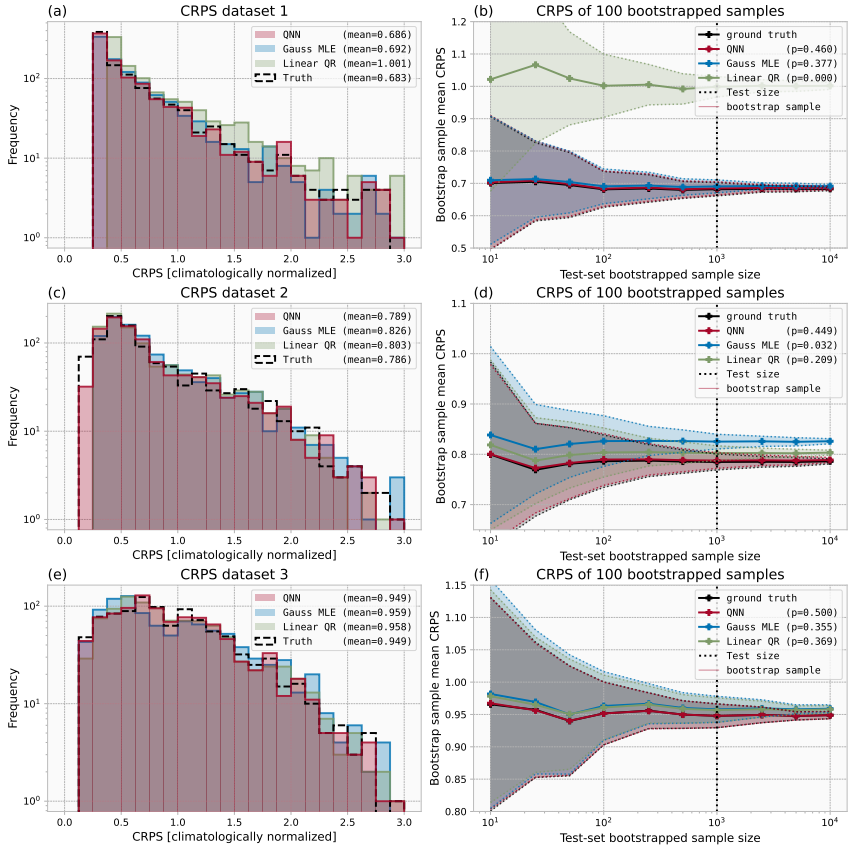

Figure 6. Continuous ranked probability score evaluated over the test samples of Synthetic Dataset 1 (a, b), Dataset 2, (c, d) and Dataset 3 (e, f). (a, c, e) Histograms of the continuous ranked probability score yielded by the predicted distributions of the RBLQNN (red), MVE network (blue), and linear quantile regression (light green) evaluated over the 1,000 test samples. The black dashed line shows the histogram of the CRPS attained by evaluating the ground truth distribution against the observed samples. The continuous ranked probability scores have been normalized relative to the climatological CRPS. Quantities in the legend indicates the sample-average CRPS for each method. (b, d, f) Bootstrap-sampled CRPS sample averages. For a variety of quasi-logspaced sample sizes N_b ranging from 10 to 10^4 (multiples of 1, 2.5, 5 times powers of 10), a bootstrapped sample of size N_b is sampled with replacement from the test set, and the CRPS sample average is taken for each method of conditional probability estimation. This is repeated 100 times to create a bootstrap ensemble. The thick notched lines show the ensemble-mean sample-average CRPS for each sample size N_b, and the shaded region shows the ensemble spread ($\pm$ one standard deviation). The black dotted line indicates the sample size of the test set ($N = 1,000$). The fractions in the legend evaluate the probability that the sample-averaged CRPS for each method will be lower than the CRPS yielded by the ground truth distribution through pairwise comparisons of bootstrapped samples with $N_b = N$. For instance, a fraction of $p = 0.03$ for the MVE network (panel d) indicates that only 3% of bootstrapped CRPS sample averages for the MVE network are lower than the CRPS sample averages computed using the ground truth distribution. This percentage is computed over the 10,000 pairwise comparisons between the 100 ensemble members for the MVE network CRPS sample averages and 100 ensemble members for the ground truth distribution CRPS sample averages.

whereas 1,000 samples are needed to reject the probabilistic predictions made by the MVE network for Dataset 2. In contrast, the probability distributions predicted by the RBLQNN cannot be rejected up to samples of size 10,000. The evaluations of the CRPS in Figure 6 indicate that even in cases in which the predicted probability distributions are superior for a given method, substantial sample sizes may be needed to discern performance using sample-based metrics such as the CRPS. Care should therefore be taken when interpreting differences between sample-mean CRPS for different methods, especially for datasets with small sample sizes.

4 Model performance on observational datasets

We next evaluate the performance of the RBLQNN against the MVE network and linear quantile regression for the GSOD daily maximum temperature datasets as well as the TRMM precipitation dataset. For the GSOD datasets, we evaluate models trained separately at 1,501 different NOAA weather stations with 1,107–1,827 test samples encompassing years 2016–2020. The TRMM dataset is a single dataset with 22,028 test samples. As the ground truth probability distribution is unknown for these datasets, performance is evaluated using sample-based metrics such as the CRPS and PIT histogram deviation statistic.

4.1 GSOD daily maximum temperatures

Figure 7 compares the sample-averaged CRPS of the RBLQNN against the various baselines. Histograms of the sample-averaged CRPS over all locations for each of the different methods are shown in Supporting Figure S6. To maintain consistency between locations with different climatological variability, CRPS values are normalized by the CRPS obtained by applying the climatological quantiles of daily temperature maximums at each location, so that a normalized sample-mean CRPS of 1 indicates probabilistic predictions no better than climatology. For all three conditional probability estimation techniques, time-mean CRPS is lower than the climatological CRPS at 1,493 out of 1,501 stations (99.4%). Moreover, the probabilistic models result in better CRPS than the MSE network at 1,492 of 1,401 locations (99.2%). Thus, all three conditional probability estimation techinques tend to provide improved information about the conditional distribution which is not permitted either by climatological or deterministic baselines.

The differences in CRPS between the RBLQNN and other conditional probabilistic baselines (Fig. 7c and 7d) are less pronounced than the difference in CRPS for the RBLQNN and climatological or MSE network predictions (Fig. 7a and 7b). Sample-mean CRPS for the RBLQNN is lower than the CRPS for the linear quantile regression at 1,477 of 1,501 locations (94.8%). The locations in which the linear quantile regression outperforms the RBLQNN could indicate overfitting on out-of-sample distributions, or random variations due to a lack of statistical significance arising from an insufficient sample size. Using pairwise comparisons of bootstrapped ensemble members to assess significance as in Section 3.4 shows that the sample-mean CRPS is statistically significantly lower for the RBLQNN than linear quantile regression at 950 of 1,501 stations (63.2%), indicating that more samples may be needed to attain statistically significant lower CRPS with the RBLQNN for many locations.

Sample-mean CRPS of the RBLQNN is lower than the MVE at 836 of 1,501 locations (55.7%). While a slight majority of locations have lower CRPS with the RBLQNN than with the MVE network, few are statistically significant (19 of 1,501 stations, 1.3%). This can indicate that conditional probability distributions of TMAX are relatively well-described using Gaussian distributions, or that more samples are needed to differentiate the skill of the RBLQNN predictions from those of the MVE network.

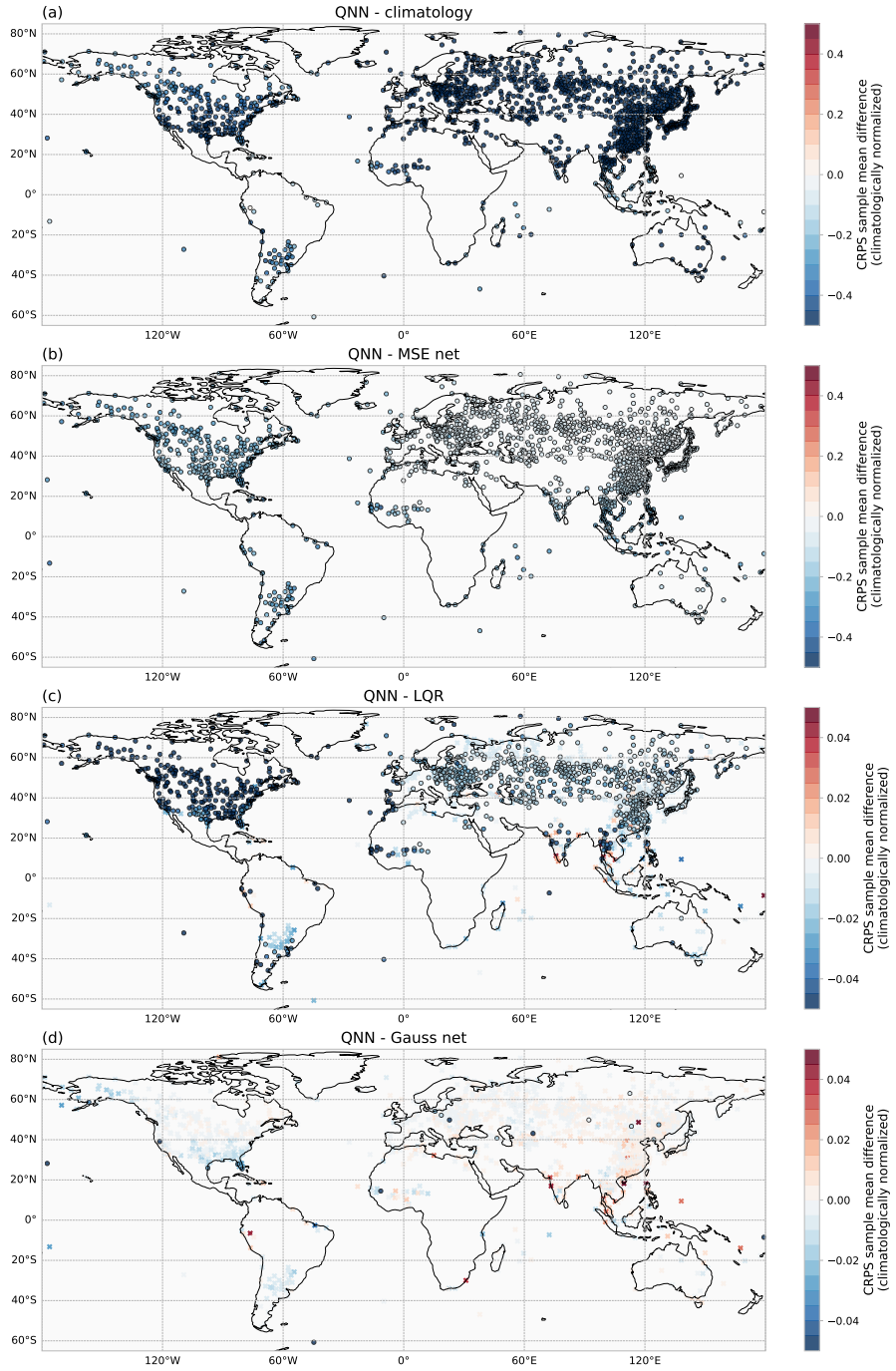

Figure 7. Comparison of sample-average CRPS between RBLQNN predictions and baselines. Maps of the difference in sample-mean CRPS between RBLQNN and (a) climatology, (b) MSE network, (c) linear quantile regression, and (d) mean-variance estimation network. Values have been normalized with respect to the climatological CRPS sample mean at each location. Circles outlined in black indicate statistically signifcant differences, determined using the pairwise comparisons method of Section 3.4. Crosses indicate differences which are not statistically significant. Note the order of magnitude difference in the colorbar extent for panels (a, b) vs (c, d).

Despite the limitations of the small test sizes for individual stations, a few cohesive geographical regions where the RBLQNN has lower sample-average CRPS than the MVE network point to systematic causes for better predictions from the RBLQNN. The greatest decrease in CRPS from using the RBLQNN instead of the MVE network occurs primarily in the southeastern United States and Alaska. PIT deviation statistics (Fig. 8) indicate that predictions are relatively well-calibrated for the RBLQNN in these regions. For instance, in North and South America, 278 of 350 stations (79%) have better calibration statistics for the RBLQNN than the MVE network. Furthermore, while 55 of 350 locations (15%) have PIT histograms fully consistent with the null hypothesis of uniformity for the RBLQNN, only 11 (3%) of the stations have PIT histograms consistent with uniformity using the MVE network. The relatively well-calibrated probabilistic predictions of the RBLQNN relative to the MVE network predictions in the southeastern United States suggest that the RBLQNN is successfully estimating inherently non-Gaussian conditional distributions of TMAX. This region is consistent with the regions of high negentropy highlighting non-Gaussian marginal distributions in Figure 1c.

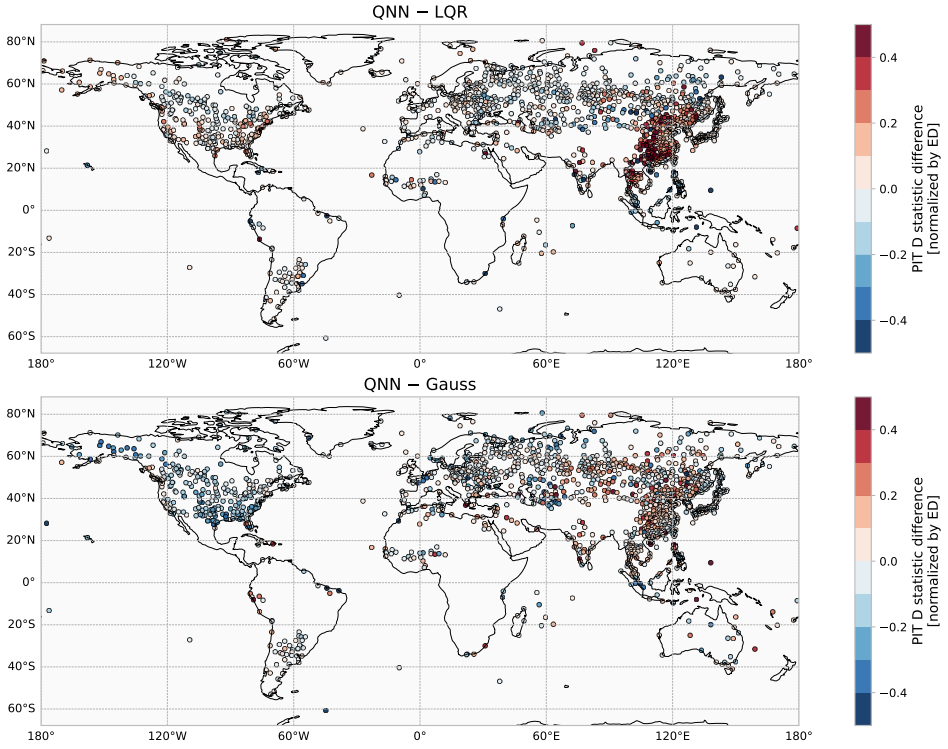

Figure 8. Differences in calibration statistics for RBLQNN predictions and those of (a) linear quantile regression and (b) MVE networks. Calibration is measured using the deviation statistic of the PIT histogram described in Section 2.3. Deviation statistics are normalized by the expected level of deviation for each location.

Regions in which the RBLQNN yields higher CRPS than the MVE network include Southeast Asia (particularly the Malay Peninsula and southeastern China), as well as Siberia and the Indian subcontinent. For these regions, the comparatively stronger performance of the MVE network indicates that the Gaussian approximation adequately represents the underlying conditional distributions. The validity of the Gaussian approximation may be related to the variability in TMAX that is attributable to functional dependence on the regressors. Supporting Figure S7 shows the distribution of differences

in CRPS between the RBLQNN and MVE network as a function of the MSE network R^2. Since the MVE network and RBLQNN share the same architecture as the MSE network, the R^2 yields an estimate about the proportion of variability which is explained by the nonlinear functional dependence of TMAX on the model inputs as opposed to the conditional distribution itself. Supporting Figure S7c suggests that the Gaussian approximation is the most valid when a high proportion of the variance is explained by the deterministic functional ($R^2 \approx 1$) or when a negligible proportion of the variance is explained by the deterministic component ($R^2 \approx 0$). In regions such as the Malay peninsula, R^2 of the MSE network is low, possibly due to the influence of the monsoon system and effects of moisture variations on the temperature profile as well as complex orography. In such regions, the inputs to the RBLQNN may be rather uninformative, resulting in challenges with optimization of individual quantiles in the RBLQNN. Conversely, in regions such as southeastern China where R^2 is high, much of the variability in TMAX is explained by the deterministic component, and the remaining variability may resemble Gaussian noise. The RBLQNN tends to outperform that of the MVE networks in regions of intermediate R^2, in which the inputs are informative yet much of the variance is not fully explained by the deterministic functional. For instance, at stations where the MSE network R^2 is between 0.3 and 0.7, 73.1% of stations have lower time-averaged CRPS with the RBLQNN than with the MVE network.

The timeseries of predicted conditional probability distributions shown in Figure 9 help illustrate the situations in which the RBLQNN performs well. Figures 9a and 9b show two stations where the RBLQNN has greater average CRPS than the MVE network. At Bangkok (Fig. 9a), temporal changes in the distribution are small over the course of the year, with the observed sea level pressure and geopotential heights providing minimal information about TMAX. On the other hand, at Fuzhou, (Fig. 9a), nearly all of the variance in temperature is explained by the deterministic component. In both of these cases, the principle of maximum entropy validates the Gaussian approximation, as there is a lack of constraining information about the conditional distributions of TMAX.

Figures 9c and 9d show two examples where the RBLQNN does outperform the MVE network, and the differences are statistically significant. The predicted probability distributions respond to seasonality, both through the change in mean predicted temperature throughout the year as well as changes in variability between the winter and summer months. However, the distributions predicted by the quantile neural network are negatively skewed, allowing the RBLQNN to permit cold extremes during the winter while maintaining sharp predicted distributions.

4.2 TRMM precipitation

While the conditional probability estimates of the RBLQNNs trained to predict temperatures in Section 4.1 mostly outperform those of linear quantile regression, differences between the RBLQNN and Gaussian maximum likelihood networks are less pronounced. This may be because the probability estimates of TMAX conditioned on sea level pressure and geopotential heights are sufficiently approximated by normal distributions at most stations, or because the test size is insufficient to assess the probabilistic predictions. In this section, we focus on a single dataset which contains significantly more samples (20,028 test samples as opposed to at most 1,827), and for which the Gaussian approximation is clearly less valid.

The CRPS is shown for the different conditional probability estimation techniques for the precipitation dataset in Figure 10a. Here, it is clear that the CRPS is substantially better for the RBLQNN than all of the other baselines. Probabilistic predictions are the worst for the Gaussian neural network, and both the Gaussian neural network and linear quantile regression produce probabilistic estimates with even worse scores than an MSE-trained deterministic network. Computing the CRPS using bootstrapped sam-

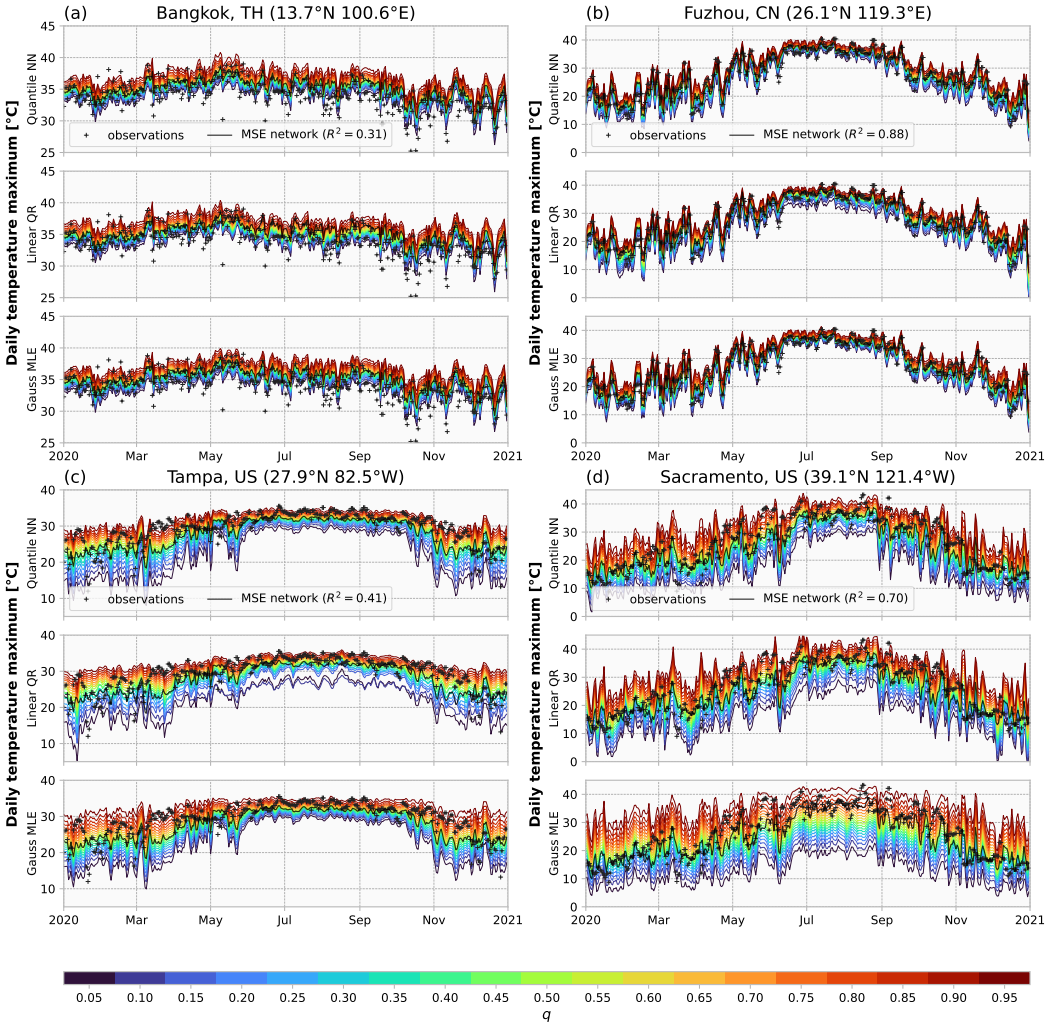

Figure 9. Timeseries of predicted conditional probability distributions for GSOD TMAX for year 2020 in (a) Bangkok, (b) Fuzhou, (c) Tampa, and (d) Sacramento. Colored lines indicate the predicted quantiles, with different panels for RBLQNN (top), linear quantile regression (middle) and mean-variance estimation networks (bottom). Black crosses indicate observed TMAX. The black line indicates predictions made by the MSE network, with R^2 given in the legend.

ples (Figure 10b) shows that for bootstrapped sample sizes of 500 samples or more, the RBLQNN consistently results in better sample-averaged CRPS than the other baselines. Thus, the RBLQNN produces significantly better probabilistic predictions on the precipitation dataset than the Gaussian maximum likelihood network or the linear quantile regression method. This indicates that both nonlinear functional dependence and non-Gaussianity are essential properties of the conditional distributions of precipitation.

5 Discussion

The chaotic and nonlinear dynamics of the Earth system implies that geophysical variables are prone to fluctuations and uncertainties, posing challenges to predicting geophysical variability with complete certainty. Thus, managing weather and climate risk requires quantifying and constraining estimates for geophysical variability which depends

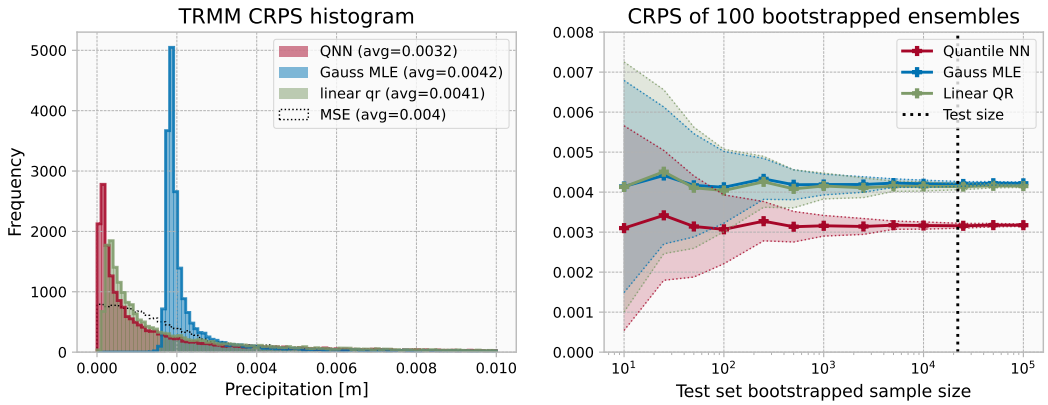

Figure 10. CRPS of different conditional probability estimation techniques evaluated on the TRMM precipitation dataset. (a) Histograms of the CRPS evaluated over all test samples for the RBLQNN (red), Gaussian maximum likelihood network (blue), linear quantile regression (light green) and MSE network (black dotted lines). Values in the legend indicate sample-averaged values. (b) As in Figure 6b, 6d, and 6f but for the TRMM dataset.

on fluctuating observable conditions. In this paper, we propose an approach for characterizing the uncertainties of geophysical quantities such as daily maximum temperatures or precipitation amounts based on other measurable conditions using quantile regression neural networks. To address some typical issues with quantile neural networks, our implementation—termed the "ReLU bias loss quantile neural network" (RBLQNN)—employs two novel and explicit modifications to the loss function to predict quantiles with equal consideration and to mitigate the possibility of predicting degenerate probability distributions. Using a suite of different datasets—synthetic distributions, in-situ daily temperature maxima observations from weather stations, and altimetry-observed precipitation data—the RBLQNN is compared against mean-variance estimation networks (which presuppose that conditional distributions are Gaussian) and linear quantile regression (where linear conditional dependence relationships are assumed). The RBLQNN is versatile, issuing conditional probability estimates which faithfully describe the target variable in the broad class of datasets considered.

We evaluate the RBLQNN on three synthetic datasets in which the true distribution is known a priori, demonstrating several minor advantages over other approaches. The RBLQNN performs well in situations where the MVE network or LQR are deficient, such as when conditional probability distributions are non-Gaussian or when the response variable depends nonlinearly on the regressors. Evaluations of the RBLQNN against other quantile neural network techniques demonstrates advantages of our approach: convergence for our method appears to be more stable than the cumulative increment approach of Padilla et al. (2022), whereas the ReLU bias loss reduces degenerate probability distributions due to quantile crossings without significantly degrading performance. Evaluations of the convergence of the RBLQNN over a large range of hyperparameters suggests that the RBLQNN trains stably over a broad hyperparameter space relative to the MVE networks. Of course, a caveat to these results is that the range of hyperparameters tested could be extended (e.g., more layers), and training may be sensitive to other hyperparameters not tested (such as activation function or optimizer).

Comparing the CRPS of the RBLQNN predictions against linear quantile regression and the MVE network predictions illustrates the relative importance of capturing nonlinearities or non-Gaussian distributions in the representation of uncertainties. For

the GSOD daily temperature maxima dataset, most stations have lower sample-averaged CRPS when using the RBLQNN than when using linear quantile regression, and comparisons between bootstrapped CRPS averages indicate that at many locations these differences are statistically significant. The relative performance of the RBLQNN over linear quantile regression implies that capturing nonlinear functional dependencies between temperature and local pressure or geopotential heights is paramount to constraining temperature uncertainties. Differences in CRPS between the RBLQNN and the MVE network are smaller and not statistically significant, indicating that allowing for non-Gaussian conditional probabilities may be of secondary importance.

In the context of daily temperature maxima, the relative validity of the Gaussian approximation in many situations may be related to the principle of maximum entropy (Sura & Hannachi, 2015; Jaynes, 1957): namely, that the distribution which maximizes the information entropy—i.e., the least informative distribution—under a set of given constraints is that which is most probable. Under the very limited constraints of given mean and variance, the Gaussian distribution maximizes the differential entropy, and thus without further information constraints the Gaussian approximation is valid. It was noted that the CRPS for the RBLQNN often was higher than the MVE networks when much of the variance was explained by the deterministic component (MSE network $R^2 \approx 1$) or when very little of the variance was explained by the deterministic component ($R^2 \approx 0$). In such cases, the Gaussian approximation may be relatively valid because the inputs are uninformative about the conditional distribution, and thus the distribution is relatively unconstrained. On the other hand, in regimes where the inputs are informative to predicting temperature, but much of the variance is left unexplained by the deterministic functional, additional information constraints may apply and the maximal entropy distribution may be more accurately described by non-Gaussian probability distributions.

While the RBLQNN does not have statistically significantly better CRPS than the MVE networks on the GSOD dataset, it is possible that the RBLQNN does predict conditional distributions of temperature more skillfully, yet that the sample size is insufficient to discern this skill. The synthetic datasets in Section 3 demonstrate that even if the RBLQNN clearly predicts the true distribution with greater accuracy (e.g. Fig. 3) a large number of samples may be needed to distinguish skill using sample based metrics like the CRPS (e.g. Fig. 6). Station temperature observations with significantly more samples may more clearly reveal the skill of the RBLQNN, though it is difficult a priori to estimate how many samples are needed for non-Gaussian statistics to emerge.

In light of the meager sample sizes for the GSOD temperature datasets and challenges identifying non-Gaussian conditional distributions, we also evaluated performance of the RBLQNN on the TRMM precipitation dataset, in which the sample size was over ten times as large and the Gaussian approximation is clearly invalid. In this case, the performance of the RBLQNN clearly outperforms the LQR and MVE network baselines. In principle, maximum-likelihood losses of other probability distributions can be used to predict parameters of different families of distributions. Since precipitation can take only nonnegative values, precipitation may be better modeled using distributions supported on the semi-infinite line, such as the exponential distribution or the Gamma distribution. Nevertheless, the RBLQNN is a simple approach to estimate conditional probabilities which does not require any assumptions about the parametric family of the underlying distribution.

Open Research Section

The NOAA Global Surface Summary of the Day dataset is available at `https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00516` (NOAA National Centers of Environmental Information, 1999). The Tropical

Rainfall Measuring Mission dataset is available at `https://gpm.nasa.gov/data/directory` (Kummerow et al., 2000). The ERA5 reanalysis dataset is available at `https://github.com/google-research/arco-era5` (Carver & Merose, 2023), courtesy of the Copernicus Climate Changes Service (C3S) Data Store (Hersbach, 2000). The code used for data processing, model training, analysis, and visualization in this study and files for replicating the software environment are provided under the MIT license at `https://github.com/andrewbrettin/quantile_ml` (Brettin, XXXX).

Acknowledgments

AB is supported by the VoLo Foundation. We thank Libby Barnes, Sara Shamekh, Carlos Fernandez-Granda, and Fabrizio Falasca for helpful discussions on this work.

References

Ahmed, F., & Schumacher, C. (2015). Convective and stratiform components of the precipitation-moisture relationship. *Geophysical Research Letters*, *42*(23), 10–453.

Ashkenazy, Y., & Smith, N. R. (2024). Data-driven analysis of annual rain distributions. *Physical Review Research*, *6*(2), 023187.

Barnes, E. A., & Barnes, R. J. (2021). Controlled Abstention Neural Networks for Identifying Skillful Predictions for Regression Problems. *Journal of Advances in Modeling Earth Systems*, *13*(12), e2021MS002575.

Barnes, E. A., Barnes, R. J., & DeMaria, M. (2023). Sinh-arcsinh-normal distributions to add uncertainty to neural network regression tasks: Applications to tropical cyclone intensity forecasts. *Environmental Data Science*, *2*, e15.

Beck, H. E., Westra, S., Tan, J., Pappenberger, F., Huffman, G. J., McVicar, T. R., ... others (2020). PPDIST, global 0.1° daily and 3-hourly precipitation probability distribution climatologies for 1979–2018. *Scientific data*, *7*(1), 302.

Belloni, A., Chernozhukov, V., Chetverikov, D., & Fernández-Val, I. (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, *213*(1), 4–29.

Bjarke, N., Barsugli, J., Hoerling, M., Quan, X.-W., & Livneh, B. (2023). When record breaking heat waves should not surprise: skewness, heavy tails and implications for risk assessment. *ESS Open Archive*.

Bocquet, M., Pires, C. A., & Wu, L. (2010). Beyond Gaussian Statistical Modeling in Geophysical Data Assimilation. *Monthly Weather Review*, *138*(8), 2997–3023.

Bourdin, D. R., Nipen, T. N., & Stull, R. B. (2014). Reliable probabilistic forecasts from an ensemble reservoir inflow forecasting system. *Water Resources Research*, *50*(4), 3108–3130.

Bremnes, J. B. (2020). Ensemble Postprocessing Using Quantile Function Regression Based on Neural Networks and Bernstein Polynomials. *Monthly Weather Review*, *148*(1), 403–414.

Brettin, A. (XXXX, XX). *Code for Brettin and Zanna (2025): Estimation of temperature and precipitation uncertainties using quantile neural networks.* Zenodo. (`https://github.com/andrewbrettin/quantile_ml`) doi: XX.xxxx/zenodo.XXXXXXXX

Bröcker, J. (2012). Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, *138*(667), 1611–1617.

Cannon, A. J. (2018). Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic environmental research and risk assessment*, *32*(11), 3207–3225.

Carver, R. W., & Merose, A. (2023). ARCO-ERA5: An Analysis-Ready Cloud-

Optimized Reanalysis Dataset. In *22nd Conf. on AI for Env. Science.* Denver, CO, USA: American Meteorological Society. (`https://ams.confex.com/ams/103ANNUAL/meetingapp.cgi/Paper/415842`)

Catalano, A., Loikith, P., & Neelin, J. (2021). Diagnosing Non-Gaussian Temperature Distribution Tails Using Back-Trajectory Analysis. *Journal of Geophysical Research: Atmospheres*, *126*(8), e2020JD033726.

Cavanaugh, N. R., & Shen, S. S. (2014). Northern Hemisphere Climatology and Trends of Statistical Moments Documented from GHCN-Daily Surface Air Temperature Station Data from 1950 to 2010. *Journal of climate*, *27*(14), 5396–5410.

Chernozhukov, V., Fernández-Val, I., & Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, *78*(3), 1093–1125.

Chronopoulos, I., Raftapostolos, A., & Kapetanios, G. (2024). Forecasting Value-at-Risk Using Deep Neural Network Quantile Regression. *Journal of Financial Econometrics*, *22*(3), 636–669.

Corsaro, S., Marino, Z., & Scognamiglio, S. (2024). Quantile mortality modelling of multiple populations via neural networks. *Insurance: Mathematics and Economics*, *116*, 114–133.

Dawid, A. P. (1982). The Well-Calibrated Bayesian. *Journal of the American Statistical Association*, *77*(379), 605–610.

Dawid, A. P. (1984). Statistical Theory: The Prequential Approach. *Journal of the Royal Statistical Society, Series A*, *147*(2), 278–290.

DelSole, T., & Tippett, M. (2022). Basic concepts in probability and statistics. In *Statistical methods for climate scientists* (p. 1–29). Cambridge University Press.

Diffenbaugh, N. S., & Barnes, E. A. (2023). Data-driven predictions of the time remaining until critical global warming thresholds are reached. *Proceedings of the National Academy of Sciences*, *120*(6), e2207183120.

Falasca, F., Brettin, A., Zanna, L., Griffies, S. M., Yin, J., & Zhao, M. (2023). Exploring the nonstationarity of coastal sea level probability distributions. *Environmental Data Science*, *2*, e16.

Fisher, R. A. (1970). Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution* (pp. 66–70). Springer.

Franzke, C. L., O'Kane, T. J., Berner, J., Williams, P. D., & Lucarini, V. (2015). Stochastic climate theory and modeling. *Wiley Interdisciplinary Reviews: Climate Change*, *6*(1), 63–78.

Garfinkel, C. I., & Harnik, N. (2017). The Non-Gaussianity and Spatial Asymmetry of Temperature Extremes Relative to the Storm Track: The Role of Horizontal Advection. *Journal of Climate*, *30*(2), 445–464.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *69*(2), 243–268.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of Atmospheric Science*, *102*(477), 359–378.

Godbole, V., Dahl, G. E., Gilmer, J., Shallue, C. J., & Nado, Z. (2023). *Deep Learning Tuning Playbook.* (Version 1.0, `http://github.com/google-research/tuning_playbook`)

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. In (chap. Chapter 8: Optimization for Training Deep Models). MIT Press. (`http://www.deeplearningbook.org`)

Gordon, E. M., & Barnes, E. A. (2022). Incorporating Uncertainty Into a Regression Neural Network Enables Identification of Decadal State-Dependent Predictability in CESM2. *Geophysical Research Letters*, *49*(15), e2022GL098635.

Guillaumin, A. P., & Zanna, L. (2021). Stochastic-Deep Learning Parameterization of Ocean Momentum Forcing. *Journal of Advances in Modeling Earth Systems*,

13(9), e2021MS002534.

Gupta, S., Mastrantonas, N., Masoller, C., & Kurths, J. (2022). Perspectives on the importance of complex systems in understanding our climate and climate change: The Nobel Prize in Physics 2021. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *32*(5).

Hassanzadeh, P., & Kuang, Z. (2015). Blocking variability: Arctic Amplification versus Arctic Oscillation. *Geophysical Research Letters*, *42*(20), 8586–8595.

Haynes, K., Lagerquist, R., McGraw, M., Musgrave, K., & Ebert-Uphoff, I. (2023). Creating and Evaluating Uncertainty Estimates with Neural Networks for Environmental-Science Applications. *Artificial Intelligence for the Earth Systems*, *2*(2), 220061.

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, *15*(5), 559–570.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... others (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049.

Hu, Y., & Pierrehumbert, R. T. (2002). The Advection–Diffusion Problem for Stratospheric Flow. Part II: Probability Distribution Function of Tracer Gradients. *Journal of the atmospheric sciences*, *59*(19), 2830–2845.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, *13*(4-5), 411–430.

IPCC. (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Tech. Rep.). Cambridge, United Kingdom and New York, NY, USA: Intergovernmental Panel on Climate Change.

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review*, *106*(4), 620.

Jiang, X., Jiang, J., & Song, X. (2012). Oracle model selection for nonlinear models based on weighted composite quantile regression. *Statistica Sinica*, 1479–1506.

Jones, C., & Pewsey, A. (2019). The sinh-arcsinh normal distribution. *Significance*, *16*(2), 6–7.

Keilbar, G., & Wang, W. (2022). Modelling systemic risk using neural network quantile regression. *Empirical Economics*, *62*(1), 93–118.

Kimura, Y., & Kraichnan, R. H. (1993). Statistics of an advected passive scalar. *Physics of Fluids A: Fluid Dynamics*, *5*(9), 2264–2277.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. doi: 10.48550/arXiv.1412.6980

Koenker, R. (2005). *Quantile regression* (Vol. 38). Cambridge university press.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, 33–50.

Kopp, R. E., Horton, R. M., Little, C. M., Mitrovica, J. X., Oppenheimer, M., Rasmussen, D., ... Tebaldi, C. (2014). Probabilistic 21st and 22nd century sea-level projections at a global network of tide-gauge sites. *Earth's future*, *2*(8), 383–406.

Kummerow, C., Simpson, J., Thiele, O., Barnes, W., Chang, A., Stocker, E., ... others (2000). The status of the Tropical Rainfall Measuring Mission (TRMM) after two years in orbit. *Journal of Applied Meteorology*, *39*(12), 1965–1982.

Laio, F., & Tamea, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences*, *11*(4), 1267–1277.

Landau, L. D., & Lifshitz, E. M. (2013). *Statistical Physics: Volume 5* (Vol. 5). Elsevier.

Li, M., Wang, G., Cao, F., Zong, S., & Chai, X. (2023). Determining optimal probability distributions for gridded precipitation data based on L-moments. *Science of The Total Environment*, *882*, 163528.

Linz, M., Chen, G., & Hu, Z. (2018). Large-Scale Atmospheric Control on Non-

Gaussian Tails of Midlatitude Temperature Distributions. *Geophysical Research Letters*, *45*(17), 9141–9149.

Loikith, P. C., & Neelin, J. D. (2015). Short-tailed temperature distributions over North America and implications for future changes in extremes. *Geophysical Research Letters*, *42*(20), 8577–8585.

Loikith, P. C., & Neelin, J. D. (2019). Non-Gaussian Cold-Side Temperature Distribution Tails and Associated Synoptic Meteorology. *Journal of Climate*, *32*(23), 8399–8414.

Lott, N., Baldwin, R., & Jones, P. (2001). *The FCC Integrated Surface Hourly Database, A New Resource of Global Climate Data* (Tech. Rep.). Asheville, NC 28801-5696: National Climatic Data Center. `https://repository.library.noaa.gov/view/noaa/13826/noaa_13826_DS1.pdf`.

Majda, A., & Wang, X. (2006). *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*. Cambridge University Press.

Martinez-Villalobos, C., & Neelin, J. D. (2019). Why Do Precipitation Intensities Tend to Follow Gamma Distributions? *Journal of the Atmospheric Sciences*, *76*(11), 3611–3631.

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*(10), 1087–1096.

McKinnon, K. A., Rhines, A., Tingley, M. P., & Huybers, P. (2016). The changing shape of Northern Hemisphere summer temperature distributions. *Journal of Geophysical Research: Atmospheres*, *121*(15), 8849–8868.

McLaughlin, R. M., & Majda, A. J. (1996). An explicit example with non-Gaussian probability distribution for nontrivial scalar mean and fluctuation. *Physics of Fluids*, *8*(2), 536–547.

Newman, R., & Noy, I. (2023). The global costs of extreme weather that are attributable to climate change. *Nature Communications*, *14*(1), 6103.

Nix, D. A., & Weigend, A. S. (1994). Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)* (Vol. 1, pp. 55–60).

NOAA National Centers of Environmental Information. (1999). *Global Surface Summary of the Day (GSOD), 1.0.* `https://www.ncei.noaa.gov/data/global-summary-of-the-day/`. (Accessed 2021-07-22)

Padilla, O. H. M., Tansey, W., & Chen, Y. (2022). Quantile regression with ReLU Networks: Estimators and minimax rates. *Journal of Machine Learning Research*, *23*(247), 1–42. (`http://jmlr.org/papers/v23/21-0309.html`)

Papacharalampous, G., Tyralis, H., Doulamis, N., & Doulamis, A. (2025). Ensemble learning for uncertainty estimation with application to the correction of satellite precipitation products. *Machine Learning: Earth*, *1*(1), 015004.

Penland, C. (1989). Random Forcing and Forecasting Using Principal Oscillation Pattern Analysis. *Monthly Weather Review*, *117*(10), 2165–2185.

Penland, C., & Sardeshmukh, P. D. (1995). The Optimal Growth of Tropical Sea Surface Temperature Anomalies. *Journal of Climate*, *8*(8), 1999–2024.

Perezhogin, P., Zanna, L., & Fernandez-Granda, C. (2023). Generative Data-Driven Approaches for Stochastic Subgrid Parameterizations in an Idealized Ocean Model. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003681.

Proistosescu, C., Rhines, A., & Huybers, P. (2016). Identification and interpretation of nonnormality in atmospheric time series. *Geophysical Research Letters*, *43*(10), 5425–5434.

Scheuerer, M., Switanek, M. B., Worsnop, R. P., & Hamill, T. M. (2020). Using Artificial Neural Networks for Generating Probabilistic Subseasonal Precipitation Forecasts over California. *Monthly Weather Review*, *148*(8), 3489–3506.

Schlick, T. (2010). *Molecular Modeling and Simulation: An Interdisciplinary Guide* (Vol. 2). Springer.

Schreck, J. S., Gagne, D. J., Becker, C., Chapman, W. E., Elmore, K., Fan, D., ... others (2024). Evidential Deep Learning: Enhancing Predictive Uncertainty Estimation for Earth System Science Applications. *Artificial Intelligence for the Earth Systems*, *3*(4), 230093.

Seitzer, M., Tavakoli, A., Antic, D., & Martius, G. (2022). On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks. In *International conference on learning representations.* (`https://openreview.net/forum?id=aPOpXlnV1T`)

Sluijterman, L., Cator, E., & Heskes, T. (2024). Optimal training of Mean Variance Estimation neural networks. *Neurocomputing*, *597*, 127929.

Smith, A., Lott, N., & Vose, R. (2011). The Integrated Surface Database: Recent Developments and Partnerships. *Bulletin of the American Meteorological Society*, *92*(6), 704–708.

Smith, L. N. (2018). *A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay.* Retrieved from `https://arxiv.org/abs/1803.09820`

Sura, P., & Hannachi, A. (2015). Perspectives of non-Gaussianity in atmospheric synoptic and low-frequency variability. *Journal of Climate*, *28*(13), 5091–5114.

Taggart, R. (2023). *Estimation of CRPS for precipitation forecasts using weighted sums of quantile scores and Brier scores.* Bureau of Meteorology.

Tambwekar, A., Maiya, A., Dhavala, S., & Saha, S. (2021). Estimation and applications of quantiles in deep binary classification. *IEEE Transactions on Artificial Intelligence*, *3*(2), 275–286.

Taylor, J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, *19*(4), 299–311.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., ... others (2017). The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bulletin of the American Meteorological Society*, *98*(1), 163–173.

Wasserman, L. (2004). *Parametric Inference* (Vol. 26). Springer.

White, H. (1992). Nonparametric estimation of conditional quantiles using neural networks. In *Computing science and statistics: Statistics of many parameters: Curves, images, spatial models* (pp. 190–199). Springer.

Wu, J., Perezhogin, P., Gagne, D. J., Reichl, B., Subramanian, A. C., Thompson, E., & Zanna, L. (2025). *Data-Driven Probabilistic Air-Sea Flux Parameterization.* (`https://arxiv.org/abs/2503.03990`)

Xu, Q., Deng, K., Jiang, C., Sun, F., & Huang, X. (2017). Composite quantile regression neural network with applications. *Expert Systems with Applications*, *76*, 129–139.

Zhang, W., Quan, H., & Srinivasan, D. (2019). An Improved Quantile Regression Neural Network for Probabilistic Load Forecasting. *IEEE Transactions on Smart Grid*, *10*(4), 4425–4434. doi: 10.1109/TSG.2018.2859749

Zhang, X., Yuan, X., Wang, C., & Song, X. (2025). Monotone composite quantile regression neural network for censored data with a cure fraction. *Computational Statistics & Data Analysis*, 108201.

Zou, H., & Yuan, M. (2008). Composite Quantile Regression and the Oracle Model Selection Theory. *Annals of Statistics*, *36*(3), 1108–1126. doi: 10.1214/07-AOS50

Supporting Information for "Nonparametric estimation of temperature and precipitation uncertainties using quantile regression neural networks"

Andrew Brettin[1]and Laure Zanna[1]

[1]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

Contents of this file

1. Figure S1: Pinball function used for optimization in the quantile neural network.
2. Figure S2: Expected quantile loss as a function of quantile probability levels.
3. Figure S3: Timeseries and histograms of samples for Synthetic Dataset 3.
4. Figure S4: Ensemble mean state vector magnitude timeseries for Synthetic Dataset 3.
5. Figure S5 Signed quantile errors on synthetic datasets as a function of the inputs.
6. Figure S6: Histograms of sample mean CRPS for all 1,501 GSOD weather stations.

Corresponding author: A. Brettin, Center for Atmosphere Ocean Science, Courant Institute of Mathematical Sciences, New York University, 251 Mercer St, New York, NY, USA. (brettin@cims.nyu.edu)

7. Figure S7: Relationship between quantile neural network skill over the Gaussian maximum likelihood network as a fucntion of the variance explained by the deterministic functional.

8. Table S1: Hyperparameter configurations for different datasets.

9. Table S2: Average number of test samples resulting in quantile crossings for different quantile neural network techniques.

10. Table S3: Proportion of ensemble members which result in no quantile crossings over all samples.

11. Table S4: Hyperparamaters varied for training stability analysis.

Introduction

Text S1.

References

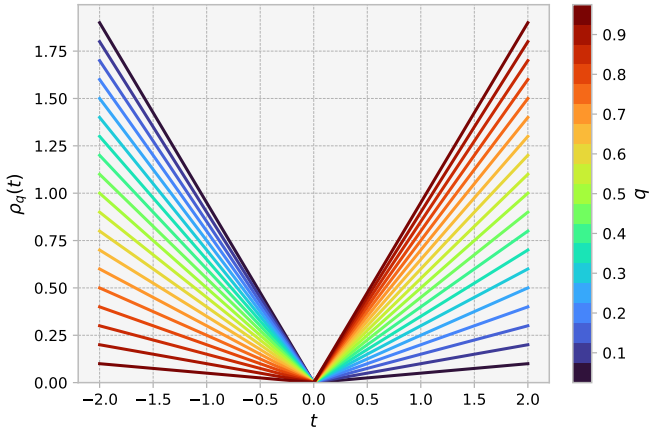

Figure S1. Pinball function used as a quantile loss (Eq. 4).

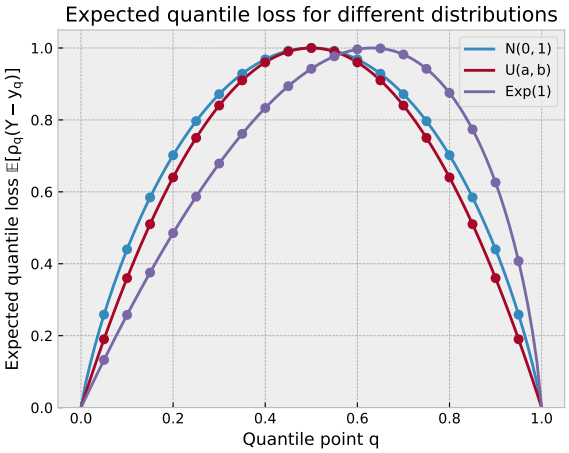

Figure S2. Expected quantile loss $\mathbb{E}\left[\rho_q(Y - y_q)\right]$ as a function of quantile probability levels q for normal (blue), uniform (red), and exponential distributions (purple).

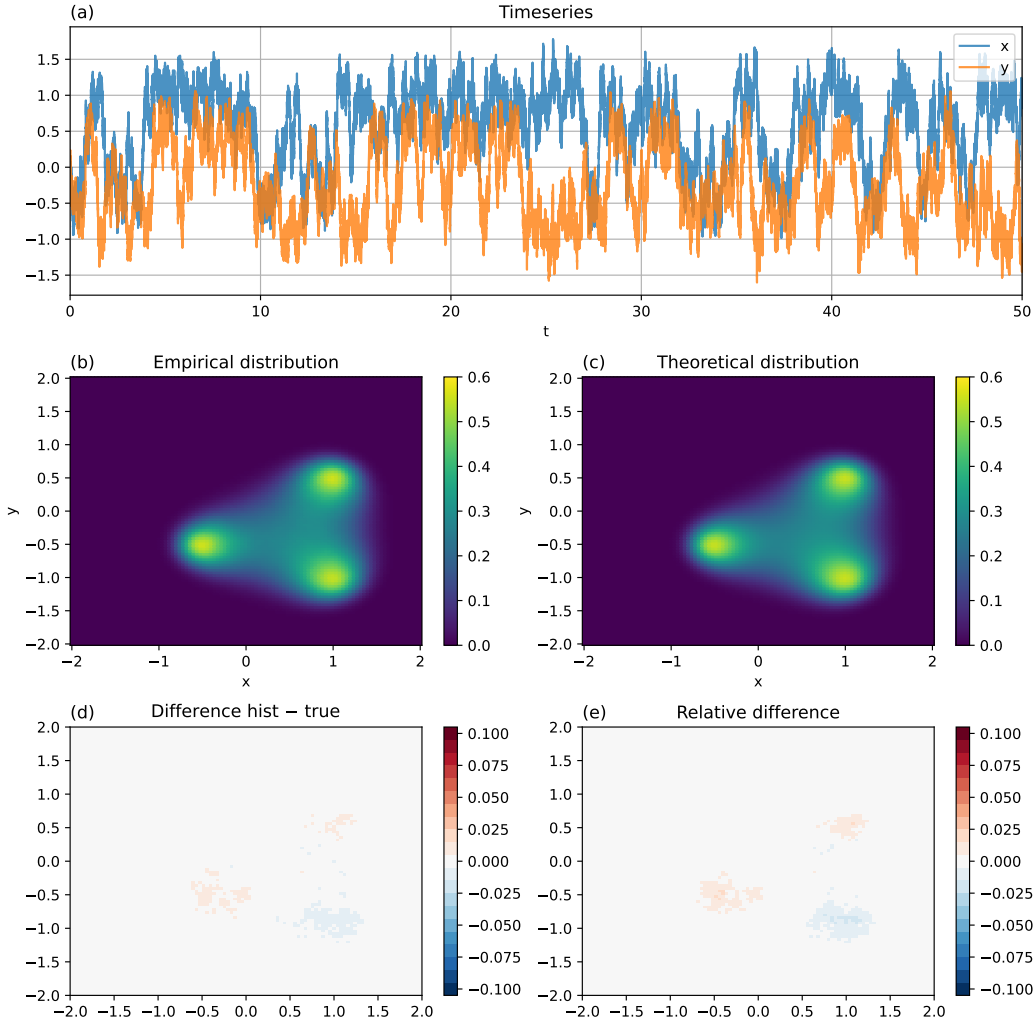

Figure S3. Timeseries and histograms of samples generated for Synthetic Dataset 3. (a) Timeseries of x (blue) and y (orange) components of the potential system (Eq. 19) for the for times $0 \leq t \leq 50$. (b) Joint histogram of samples of (x, y) for times $0 \leq t \leq 10,000$. (c) Theoretical stationary distribution given by the Fokker-Planck equation (the Boltzmann distribution $e^{-V(x,y)}$) for the potential given in Eq. 20. (d) Difference between the empirical histogram (c) and true density (d), $\tilde{f} - f$. (e) Relative difference between the histogram and true density $(\tilde{f} - f)/f$.

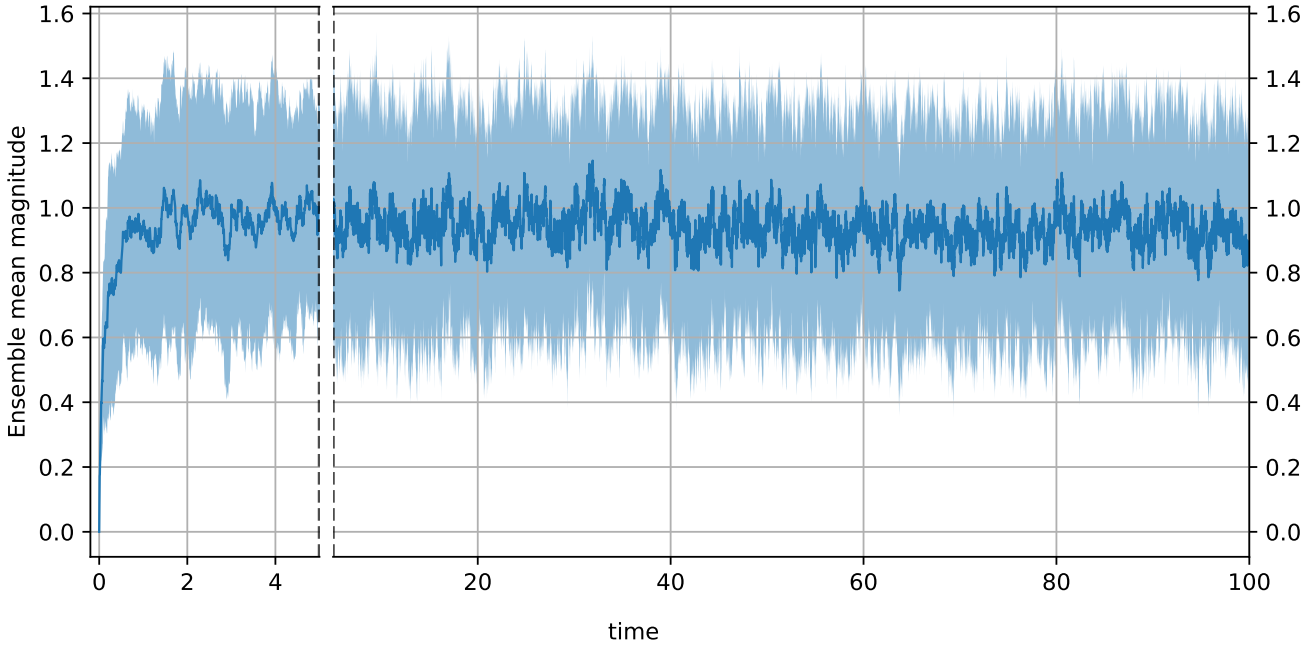

Figure S4. Ensemble mean position vector magnitude $\|(x, y)\|_2$ over 50 different trajectories for the first 100 seconds of simulation time. Shading indicates the ensemble spread ($\pm 1\sigma$).

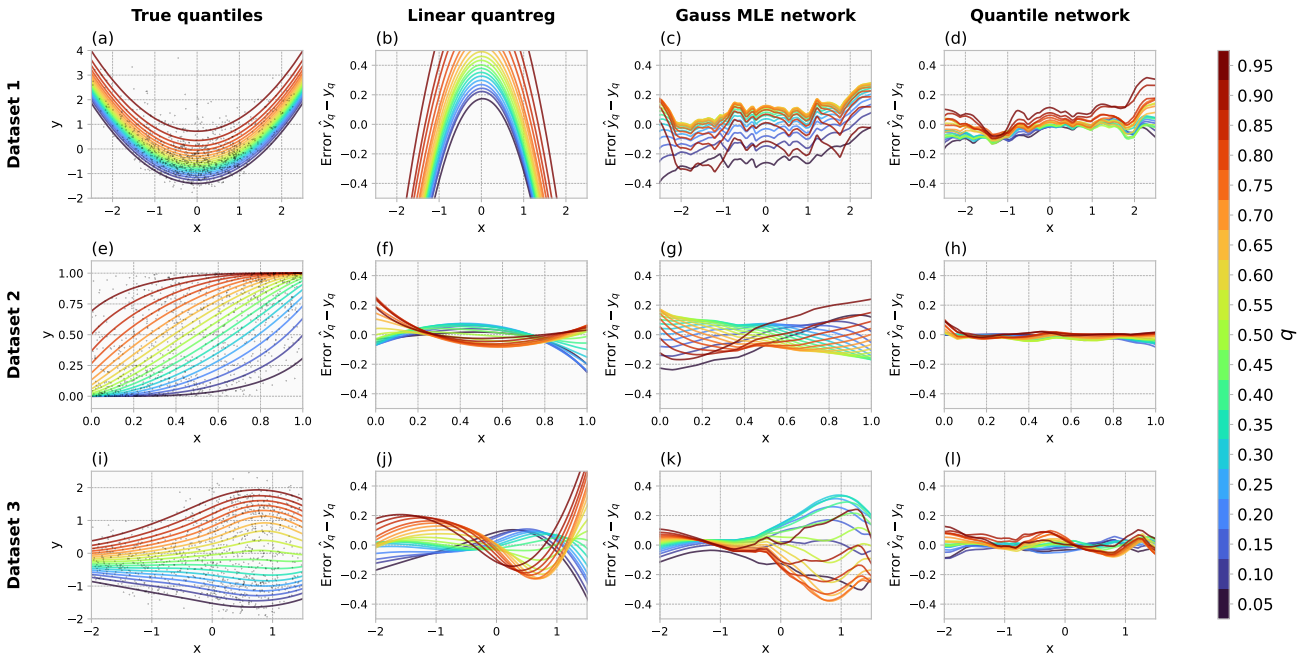

Figure S5. Quantile errors $\hat{y}_q - y_q$ as a function of x.

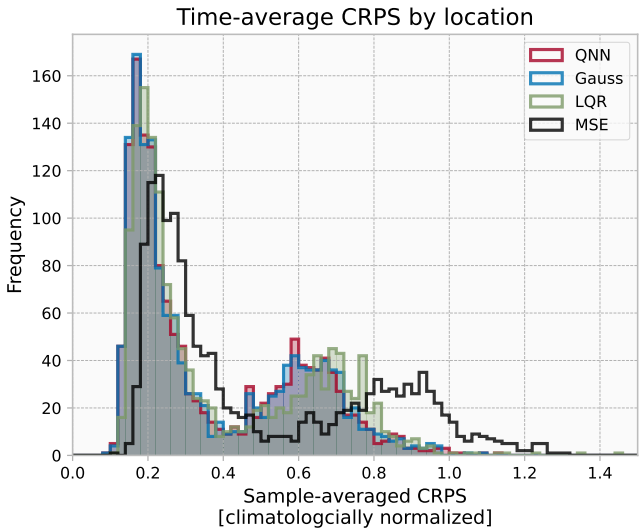

Figure S6. Histogram of sample-mean CRPS over all 1,501 GSOD stations for quantile neural network (red), Gaussian maximum likelihood network (blue), linear quantile regression (green), and MSE network (black). Values have been standardized with respect to the climatological CRPS at each location.

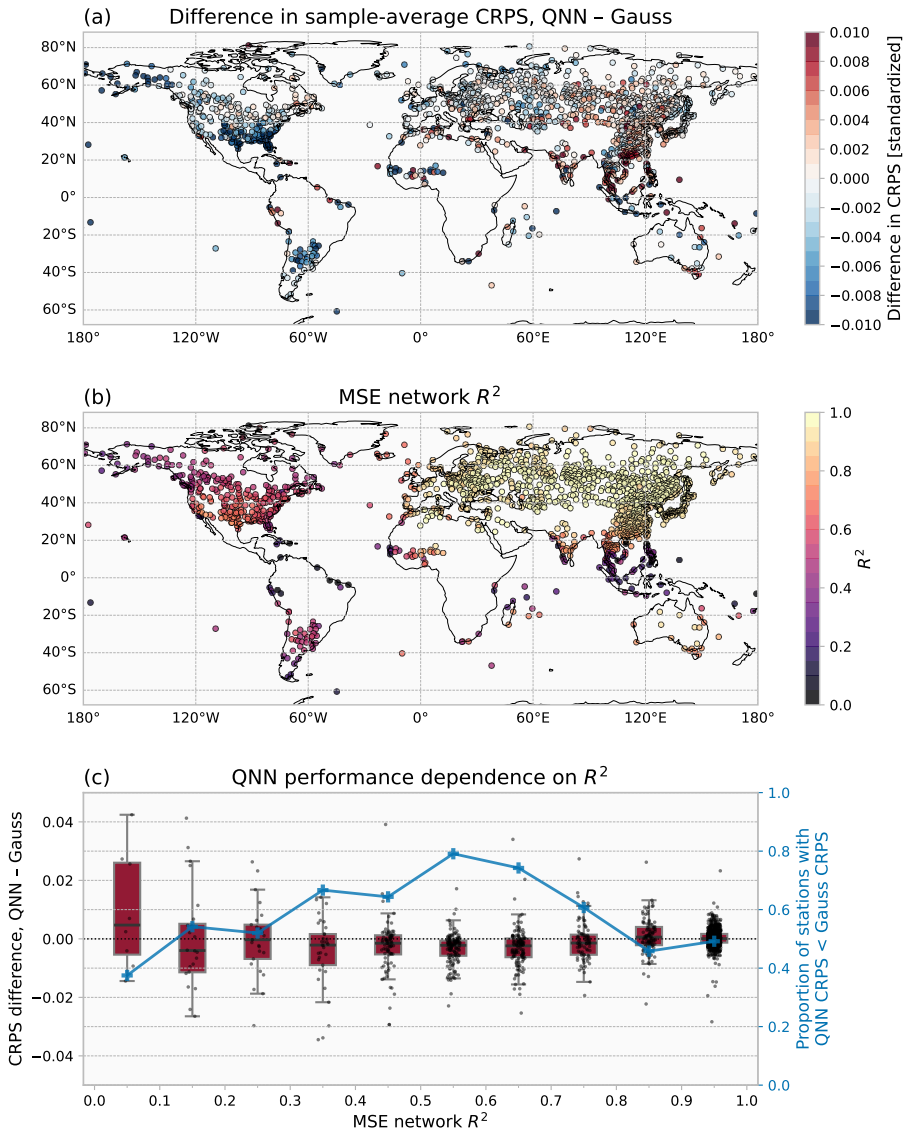

Figure S7. Relationship between CRPS gain from Gaussian maximum likelihood network to quantile neural network and coefficient of determination for MSE network. (a) Sample-average CRPS difference between quantile neural network and Gaussian maximum likelihood network, as in Figure 7d. (b) Coefficient of determination (R^2) for MSE network. (c) Boxplots of CRPS difference conditioned on corresponding station MSE network R^2 (binned into tenths), with strip plots (black dots) indicating individual stations. Blue line indicates the proportion of stations with lower CRPS for the quantile neural network.

Table S1. Hyperparameter configurations for different datasets.

Hyperparameter	Synthetic datasets	GSOD dataset	TRMM dataset
Hidden layers	3	3	3
Neurons per hidden layer	128	128	128
Bias loss weight η (QNN)	1.0	0.01	0.01
Batch size	256	256	256
Optimizer	Adam	Adam	Adam
Learning rate	10^{-3}	10^{-5}	10^{-5}
L_2 regularization	10^{-5}	10^{-2}	10^{-2}
Maximum epochs	1,000	1,000	1,000
Early stopping epochs	25	300	None
Warmup epochs	100	100	100
Prescribed variance	0.1	0.1	0.1

Table S2. Number of test samples (out of 1,000) resulting in quantile crossings for each quantile neural network technique, averaged over all 100 ensemble members.

Dataset	Quantile NN	Unweighted	No bias	Cumulative inc.
1	2.1	6.06	5.5	0.0
2	26.5	266.2	258.8	0.0
3	0.0	0.50	1.1	0.0

Table S3. Proportion of ensemble members resulting in no quantile crossings over all samples.

Dataset	Quantile NN	Unweighted	No bias	Cumulative inc.
1	0.38	0.02	0.05	1.0
2	0.48	0.00	0.00	1.0
3	0.98	0.79	0.76	1.0

Table S4. Hyperparameters used for training stability analysis.

Hyperparameter	Values considered							
Learning rate	10^{-8}	10^{-7}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
L_2 weight decay	0	10^{-5}	10^{-3}					
Number of layers	2	3						
Neurons per layer	64	128	256					