Environmental Data Science (2025), 4: e26, 1–13
doi:10.1017/eds.2025.16

METHODS PAPER

Reflective error: a metric for assessing predictive performance at extreme events

Robert Edwin Rouse[1,2], Henry Moss[2], Scott Hosking[3,4], Allan McRobie[1] and Emily Shuckburgh[5]

[1]Department of Engineering, University of Cambridge, Cambridge, UK
[2]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK
[3]AI Lab at British Antarctic Survey, Cambridge, UK
[4]The Alan Turing Institute, London, UK
[5]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
Corresponding author: Robert Edwin Rouse; Email: rer44@cam.ac.uk

Received: 07 December 2023; **Revised:** 16 March 2025; **Accepted:** 04 April 2025

Keywords: error metrics; extreme values; machine learning; natural hazards; statistics

Abstract

When using machine learning to model environmental systems, it is often a model's ability to predict extreme behaviors that yields the highest practical value to policy makers. However, most existing error metrics used to evaluate the performance of environmental machine learning models weigh error equally across test data. Thus, routine performance is prioritized over a model's ability to robustly quantify extreme behaviors. In this work, we present a new error metric, termed **Reflective Error**, which quantifies the degree at which our model error is distributed around our extremes, in contrast to existing model evaluation methods that aggregate error over all events. The suitability of our proposed metric is demonstrated on a real-world hydrological modeling problem, where extreme values are of particular concern.

Impact Statement

This paper addresses the lack of suitable metrics for assessing model performance at extreme events. We aim to develop a method that can identify whether the source of error is being driven by poor performance around extremes or performance on routine data. The authors propose a weighting function, derived from the observed data, to create a metric that enables practitioners to quantify this; the metric's utility is demonstrated in a standard hydrology problem. The authors hope that this metric can facilitate the design and identification of machine learning models better at predicting extreme events, such as flooding, storms, or heatwaves.

1. Introduction

1.1. Background

Often in real world problems, the main body of data is less interesting than the outliers, whilst those extremes are of far more concern; meteorological phenomena, such as heatwaves, storms, or floods, can result in loss of human life, displacement of local populations, destruction of built environment infrastructure, and considerable economic disruption (Barriopedro et al., 2011; Changnon et al., 2000;

This research article was awarded Open Materials badge for transparent practices. See the Data Availability Statement for details.

Easterling, 2000; Fouillet et al., 2006; Jonkman et al., 2009). In the field of hydrology, for example, extensive flooding throughout the UK in 2012 and 2014, arising from successive extreme weather systems and subsequent rainfall, caused significant damage to infrastructure, with total economic damages estimated at approximately £0.6 billion and £1.3 billion, respectively (Parry et al., 2013; The Environment Agency, 2013; Muchan et al., 2015; Chatterton et al., 2016). In 2012 alone, approximately 20% of the days of the year were classed as flood days. Being able to predict the hydrological behavior in such events is critical for developing adaptation and mitigation strategies. To be able to develop hydrological models capable of predicting these singularities with a high degree of accuracy, we require robust tools to quantitatively assess their performance with regard to said singularities or extremes. However, many of the metrics used to assess and train machine learning models do not reflect the relative importance of these extremes.

For regression tasks, we commonly see metrics such as the root mean squared error (RMSE) or R^2 (Hastie et al., 2009). However, these metrics, whilst sensitive to the magnitude of discrepancy, are insensitive to the location of errors within the target domain (Dawson et al., 2006). Of course, one could choose to ignore certain subsets of the data and instead build metrics that are more sensitive to subsets of the target domain, for example, by applying a relevance mapping based on statistical thresholds (Torgo and Ribeiro, 2007; Ribeiro and Moniz, 2020), though the selection of statistical thresholds might be considered somewhat arbitrary.

The field of extreme value theory, which is concerned with the asymptotic distribution of maxima (Haan and Ferreira, 2006), has been used to develop modeling frameworks for predicting extreme events (Ding et al., 2019; Siffer et al., 2017; Boulaguiem et al., 2022). This research includes the development of an extreme loss function based on the Fisher–Tippett–Gnedenko theorem for neural models (Ding et al., 2019) and outlier detection using automatically setting thresholds (Siffer et al., 2017). Although these methods allow for evaluating model performance on extreme values, our objective is to develop a framework that does not require event separation and that can be applied across the target domain more generally, covering local minima and maxima as well, such as those for Gaussian mixture distributions.

In this work, we develop a dimensionless metric for assessing the distribution of errors in the target domain more generally. This metric being dimensionless enables comparison between different instances in the same field. If we again use the hydrological context as a motivating example, we might wish to apply the same modeling approach to different rivers. However, their average discharge values will likely be different and, thus, the unnormalized mean error arising from model predictions will be of different scales. Our metric enables quantification of relative performance around extremes compared with what we term **routine data**, being the data that are considered close to the peaks of probability density, with respect to the probability distribution that best fits the observed data. Our hope is that doing so can supplement analysis and enable the evaluation of model performance according to the distribution of the observations with a view to more targeted model optimization. We begin by reintroducing a commonly used error metric and from it derive a new error metric. We then provide synthetic, illustrative data to which we apply this metric along with a real-world application, one where extremes are of considerable importance.

2. Methodology

2.1. Root mean squared error

Our proposed metric is derived from the ubiquitous RMSE. RMSE is expressed as the sum of the error calculated between all pairwise observations, y_i, and predictions, y_i', as in Equation 2.1; in a machine learning context, it is typically evaluated both during training, giving performance for the fit over the training set, and when testing a model to evaluate its ability to generalize (Hastie et al., 2009).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - y_i'\right)^2}{n}} \tag{2.1}$$

https://doi.org/10.1017/eds.2025.16 Published online by Cambridge University Press

The RMSE lies within the interval $[0, \infty)$, where 0 is a perfect score, and higher scores indicate lower model performance. Whilst this metric provides an overall assessment of the accuracy of a model, it has two flaws: (1) RMSE's lack of variance to data rescaling means that it does not allow for comparison of a single model archetype across application domains where the observations are of different scales; and (2) RMSE does not provide sensitivity to the distribution of errors. The implication of the latter issue, which is common to other error metrics such as R^2, is that these metrics cannot be used to identify whether or not extremes are the main contributor to the error. For the real-world problems identified in Section 1, being able to assess predictive power around extremes is essential.

2.2. Reflective error

We define our reflective error (RE) metric as the ratio between a weighted form of RMSE, such as that proposed by (Ribeiro and Moniz, 2020), and the standard RMSE. It alleviates both of the issues facing RMSE because (1) dividing the weighted form of RMSE by the standard form of RMSE is a normalization process that places all applications of RE onto the same scale and (2) the weighting function we utilize is derived by fitting an empirical distribution to the target outputs and, thus, provides local sensitivity by minimizing routine error relative to extreme error. For machine learning applications, this would be fitted to the training data targets. Throughout this paper, we refer to this process as fitting the **underlying probability distribution**, $U(y)$. From this fitted probability distribution, $U(y)$, we define the reflective weighting function, $\Psi(y)$, in Equation 2.2, with scaling factor $\kappa = \max_y(U(y))$.

$$\Psi(y) = -\alpha \cdot \frac{U(y)}{\kappa} + \beta \tag{2.2}$$

Where $\alpha = 1$ and $\beta = 1$, such that the weighting function is applied to the error at any given y and is a mapping $\Psi(y) : Y \rightarrow [0,1] \forall\, y \in Y$. Consequently, $\Psi(y) \rightarrow 1$ for extremes and $\Psi(y) \rightarrow 0$ for routine data. Thus, we formulate our metric in terms of $\Psi(y)$ and the squared error in Equation 2.3:

$$RE = \left(\frac{\sum_{i=1}^{n} \left(y_i - y_i'\right)^2 \cdot \Psi(y_i)}{\sum_{i=1}^{n} \left(y_i - y_i'\right)^2} \right)^{\frac{1}{2}} \tag{2.3}$$

For most applications, RE lies within the interval $[0, 1]$, though for instances where the total error is 0 or ∞ RE is undefined. Where RE is defined, 0 would indicate that all error comes from the point of highest probability density and 1 would indicate that all error comes from the most extreme data point; whilst these scores are unlikely to occur in real problems, we still intuit that high RE indicates the error is primarily extreme driven and low RE that the error is routine driven. This stationary expression of RE is applied to a fictitious dataset in Section 2.3.

Should the practitioner be dealing with a distribution with multiple tails, but where only one is of interest or they need separate treatment, then the weighting could be split piecewise about the global maximum probability density or other local maxima. For example, if considering a two-tailed distribution, such as a Gaussian, and the higher extremes were of more concern, then the following weighting in Equation 2.4 could be applied.

$$\Psi_{S^+}(y_i) \sim \begin{cases} 0 & \text{if } y_i \leq \underset{y}{argmax}(U(y)) \\ \Psi(y_i) & \text{if } y_i > \underset{y}{argmax}(U(y)) \end{cases} \tag{2.4}$$

Where $\Psi(y)$ is the base form of our metric, as in Equation 2.3, and Ψ_{S^+} is the piecewise form.

2.3. Stationary synthetic data

In order to demonstrate how the sensitivity of RE to extreme versus routine errors, we present a one-dimensional problem, using a normal distribution, to which we fit a model with some error. We then

https://doi.org/10.1017/eds.2025.16 Published online by Cambridge University Press

proceed to magnify that error at routine datapoints and extremes to show how RE varies whilst RMSE and R^2 are relatively unaffected.

We create an artificial, normally distributed dataset $X \sim \mathcal{N}\left(3.5,0.75^2\right)$ of 500 data points, which is of similar size to datasets commonly used to demonstrate statistical methods, such as the Iris petal or Ionosphere datasets Sigillito et al. (1989); Bezdek et al. (1999). This dataset will form our observations, and our predictive "model" is the same data with the addition of some Gaussian noise term $\epsilon_1 \sim \mathcal{N}\left(0,0.1^2\right)$. This model fits the data with $RMSE = 0.203$, $R^2 = 0.924$, and $RE = 0.542$.

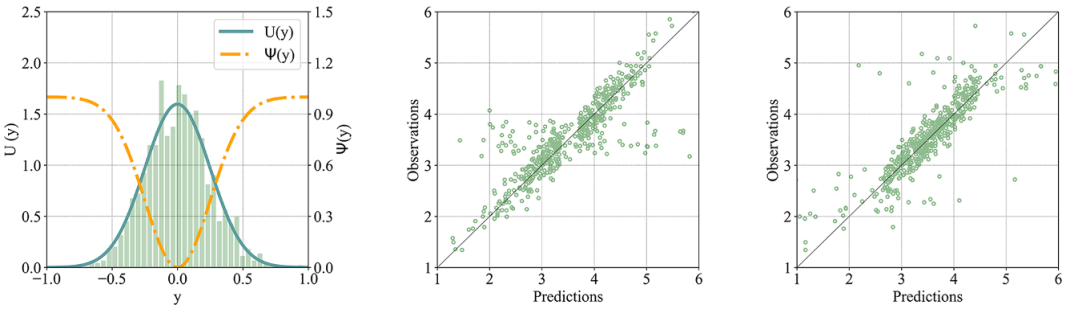
For both of the error scenarios, we add some randomly generated error, $\epsilon_2 \sim \mathcal{N}\left(0,1.0^2\right)$. For our routine error scenario, we add this error to the 100 data points closest to the mean, whilst for the extreme error scenario, we add this to the 100 data points furthest from the mean, in either direction. The predictions against observations for both of these scenarios, along with a histogram of the original dataset and accompanying relative weighting function, $\Psi(y)$, is shown in Figure 1.

For the routine error scenario, the $RMSE = 0.48$, $R^2 = 0.60$, and $RE = 0.31$; for the extreme error scenario, the $RMSE = 0.49$, $R^2 = 0.59$, and $RE = 0.80$. The values of the standard error metrics we used, the RMSE and R^2, for both of the scenarios are roughly equivalent. However, there is a significant discrepancy between the RE values; a high value of RE for the extreme scenario indicates the error is skewed towards extremes, and the opposite is true for the routine error scenario. Therefore, our RE metric is sensitive to the location of error and is identifying the corresponding relative contribution of error as intended.

3. Nonstationary methodology

3.1. Nonstationary RE

Our initial definition of RE holds assuming that the $U(y)$ is stationary and does not change over time; however, for some practical problems, the stationary assumption is not appropriate and the above, stationary form of RE does not immediately apply. For example, if we consider the trend in mean global surface temperature over the past 150 years, then what was extreme in 1850 would not be considered extreme today (V. Masson-Delmotte et al., 2021); thus, our definition of extreme must change over time and so too the weighting we apply. We therefore extend our metric to enable application to nonstationary problems, where the probability distribution of the target domain data and the associated parameters change over time, and we require local sensitivity.



(a) Normal distribution histogram with fitted probability density and reflective weighting functions

(b) Perturbed routine predictions versus observations

(c) Perturbed extreme predictions versus observations

Figure 1. *Normally distributed dataset of synthetic observations with fitted probability density and reflective weighting functions (left) with the predictions versus observations for the two perturbation scenarios (center and right).*

https://doi.org/10.1017/eds.2025.16 Published online by Cambridge University Press

If we take some subdomain, ϕ, of the whole target domain, Y, such that the subdomain could be a series of points between two limits or at a single point in time, then we can determine a subdomain-specific weighting function, $\Psi_\phi(y)$. For a range of points between two limits, such as for a step change, we would empirically fit a probability distribution, $U_\phi(y)$, to the data between those limits; if, however, there were a trend in the probability distribution, then we could empirically derive parameters of the probability distribution at a given point according to some binning strategy. This weighting function, $\Psi_\phi(y)$, is then expressed in terms of the subdomain's probability distribution, $U_\phi(y)$, as in Equation 2.2, with a scaling factor $\kappa = \max_y \left(U_\phi(y)\right)$.

$$RE = \left(\frac{\sum_{i=1}^{n} \left(y_i - y_i'\right)^2 \cdot \Psi_\phi(y_i)}{\sum_{i=1}^{n} \left(y_i - y_i'\right)^2} \right)^{\frac{1}{2}} \tag{3.1}$$

Where $\Psi_\phi(y_i)$ is the weighting to be applied to the error at any given y based on the underlying, probability distribution corresponding to values y within the local subdomain, ϕ, of the target domain, Y. More concretely, for any $\phi \subseteq Y$, we can define a weighting function $\Psi_\phi(y) : \phi[0,1]$. Note that for stationary cases, where the subdomain $\phi = Y$, this simplifies to the form of RE, expressed in Equation 2.3.

For the set of subdomains, Φ, defining the temporal limits $\forall\phi \in \Phi$ impacts the resulting probability distributions, $\Psi_\phi(y)$, and care must therefore be taken on the identification of the evolution of nonstationary trends over time. Methods for establishing nonstationary trends (Bell, 1984; Box and Tiao, 1965; Wu et al., 2007) can be used to establish the nonstationarity of statistical parameters and guide the construction of Φ. We also note that characteristic information or extrema may be missing from undersampled or poorly aliased data (Proakis and Manolakis, 2007), for example, those not complying with the Nyquist rate; in such cases, the probabilistic distribution to be fitted to any subdomain, ϕ, would likely not be representative of the true signal and the application of nonstationary RE difficult.

3.2. Nonstationary fictitious data

To illustrate the need for local sensitivity, for cases where $\phi_n \subset Y$, we present a fictitious, nonstationary signal with step changes, requiring the empirical derivation of the probability distribution before and after each step change. These observations are generated according to the piece-wise function, in Equation 3.2.

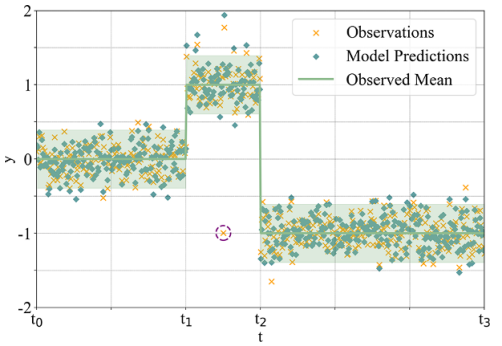
$$U(y) \sim \begin{cases} \mathcal{N}\left(0, 0.2^2\right) & \text{if } t_0 \leq t < t_1 \\ \mathcal{N}\left(1, 0.2^2\right) & \text{if } t_1 \leq t < t_2 \\ \mathcal{N}\left(-1, 0.2^2\right) & \text{if } t_2 \leq t < t_3 \end{cases} \tag{3.2}$$

The probability distribution over the entire domain temporal domain is, therefore, given by a Gaussian mixture distribution. Similar to the process for the simple stationary experiment, we create our model by adding some Gaussian error to the observations, $\epsilon \sim \mathcal{N}\left(0, 0.1^2\right)$. We then take a single prediction from the second subdomain, where $t_1 \leq t < t_2$ and $\mu_2 = 1$, and add significant error to it such that it lies close to the mean, $\mu_3 = -1$, of the third subdomain, $t_2 \leq t < t_3$, as shown in Figure 2. We will term this anomalous data point y_δ. If we were to use the stationary form of RE, then given that this error lies close to the maximum, $\Psi(y_\delta) \to 0$; however, the nonstationary form of RE is such that $\Psi(y_\delta) \to 1$ and the effect of y_δ is maximized. The resulting $RE = 0.45$ if we assumed stationarity but, by accounting for temporal variation in the fitted probability distributions, $RE = 0.83$, with the metric showing more skew towards errors around the extremes. Obviously, this dataset is fictitious but it does demonstrate a potentially desirable change in behavior in the face of nonstationarity.

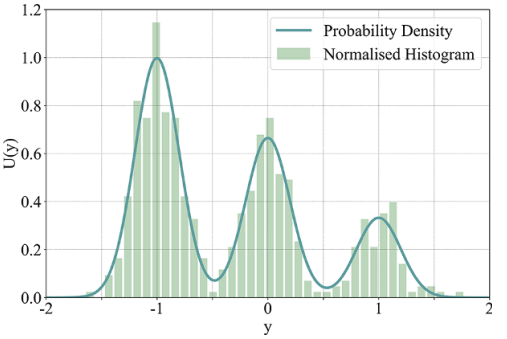
4. Application

As a real-world example, we present a hydrological modeling problem, using machine learning, where the extremes are of particular concern, given that these extremes have the potential to give rise to significant

https://doi.org/10.1017/eds.2025.16 Published online by Cambridge University Press



(a) Fictitious nonstationary data series with identifiable subdomains and anomalous data point marked



(b) Full domain probability density function with normalised histogram of fictitious observations

Figure 2. *Synthetic temporally variant dataset with predictions and observations shown about the mean function and within 2 standard deviations with anomalous observation highlighted (left); and the overall probability distribution fitted to the data with normalized histogram of the observations (right).*

flooding Faulkner (2008); Kendon and McCarthy (2015); Muchan et al. (2015). The use and potential of machine learning in hydrology has been covered in literature Abrahart et al. (2012); Besaw et al. (2010); Govindaraju et al. (2000), so we assume its application is likewise suitable here. In addition to RMSE, we will also use Nash-Sutcliffe efficiency (NSE) Nash and Sutcliffe (1970); mathematically similar to R^2, NSE is a normalization that enables comparison across catchments. Whilst NSE and other related metrics, such as Kling-Gupta Efficiency Gupta et al. (2009), are often used in conjunction with RMSE Ritter and Muñoz-Carpena (2013), our metric will supplement this further through the attribution of error to extremes, as mentioned at the beginning of Section 2.

Using (Rouse et al., 2025) as the experimental basis, we take mean gauged daily streamflow observations for a pair of river catchments, the River Avon at Bathford and the River Exe at Pixton, from the United Kingdom's National River Flow Archive (NRFA) provided by the UK Centre for Ecology & Hydrology UK Centre for Ecology & Hydrology (2022). These two rivers have been selected from the database due to the relatively different modeling challenges they present, in that the Exe at Pixton is a much smaller, flashier catchment, with an area of 159.7 km^2 compared to 1552 km^2 for the Avon at Bathford, and will thus likely make the prediction of extremes harder. This is exacerbated by the very low permeability of the catchment, which is shown amongst other characteristics in Figure 3.

Input meteorological data has been taken from ERA5, the fifth generation of global climate reanalysis modeling and data assimilation output produced by The European Centre for Medium-Range Weather Forecast's (ECMWF), which we assume to be congruent with observations Hersbach et al. (2023, 2020); Tarek et al. (2020); in spite of this, we note that the resolution of ERA5 data, at a grid-scale of 31 km, may not fully capture the spatial dynamics driving flow in a small catchment like the Exe at Pixton but that downscaling the precipitation is out of scope for this work. We use 14 days' worth of surface precipitation and daily average temperature, relative humidity, and resultant wind speed at a pressure level of 1000 hPa within the catchment along with antecedent proxies for soil moisture, as per (Rouse et al., 2025). Similarly, we adopt the model setup and training procedure for a simple artificial neural network Rumelhart et al. (1986), with layer nodes of $56 \rightarrow 16 \rightarrow 4 \rightarrow 1$, Sigmoid Linear Unit activation function Ramachandran et al. (2017), and the Adam optimization algorithm Kingma and Ba (2017). The data, running from 1979 to 2019, is split into the same test, validation, and training subsets, with the training set containing the years 1979–2008 and the test set containing the years 2011–2019.

We have assumed that the streamflow, Y, is simulated by a lognormal distribution with parameters, μ and σ, such that $Y \sim \mathcal{LN}(\mu, \sigma^2)$. Using maximum likelihood estimation (MLE) fitted to a 365-day rolling window of streamflow, we determined that μ and σ did not exhibit any significant temporal trend over the

https://doi.org/10.1017/eds.2025.16 Published online by Cambridge University Press

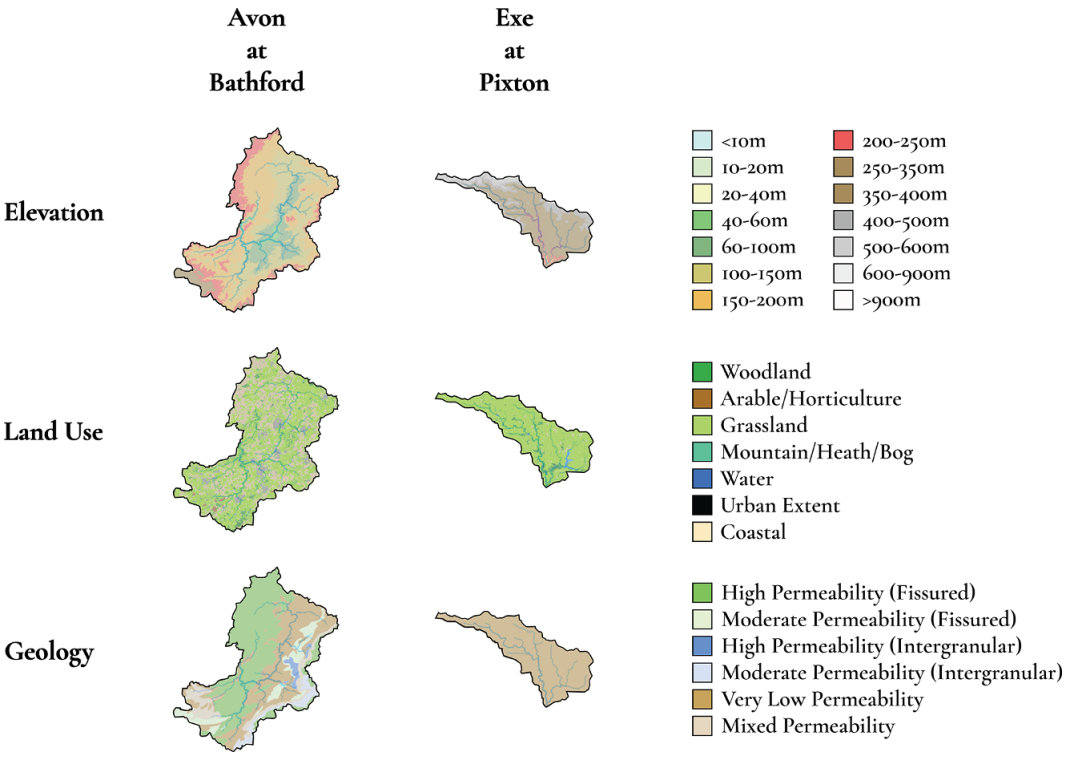
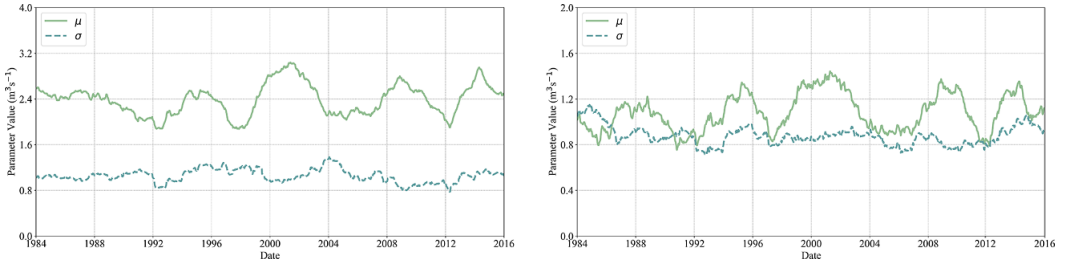


Figure 3. *Maps of elevation, land use, and geology for the Avon at Bathford and Exe at Pixton (note that the catchments are shown at different scales for visibility). Keys corresponding to each map represent the proportion of each within respective subcategories. Adapted from the National River Flow Archive UK Centre for Ecology & Hydrology (2022).*



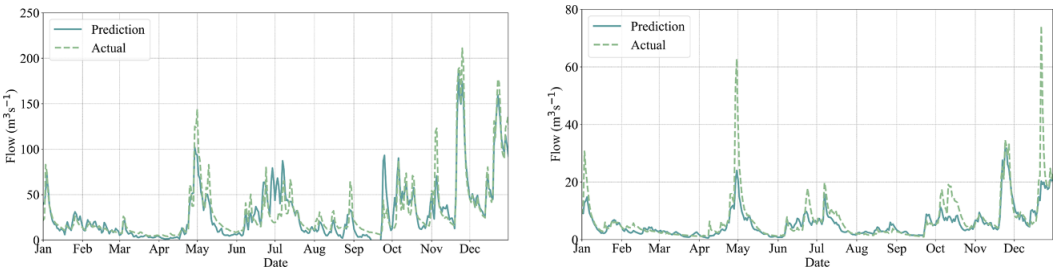
(a) The River Avon, gauged at Bathford

(b) The River Exe, gauged at Pixton

Figure 4. *Parameters for a log-normal probability density function fitted to a 365-day rolling window of streamflow observations for the River Avon and the River Exe.*

study period, as shown in Figure 4. Therefore, although climate is nonstationary, we believe the stationary assumption to be valid, and so we use stationary RE. For problems with clear nonstationarity, we would use the nonstationary formulation of Equation 3.1 and, given that we would expect a gradual rather than step change in statistical distribution, would apply a rolling window of appropriate timescale such that the weighting function, $\Psi_\phi(y_i)$, is likewise continuous. The 365-day rolling window as used above window would likely capture seasonality but could be set longer to capture longer-term climatic oscillations, such as the El Niño–Southern Oscillation, if pertinent.

https://doi.org/10.1017/eds.2025.16 Published online by Cambridge University Press



(a) The River Avon, gauged at Bathford

(b) The River Exe, gauged at Pixton

Figure 5. *Predicted and observed streamflow time series in the year 2012 for the River Avon and River Exe, with predictions generated using a basic artificial neural network model.*

Test set performance metrics are: $RMSE = 9.35 m^3 s^{-1}$, $NSE = 0.87$, and $RE = 0.92$ for the River Avon at Bathford; and $RMSE = 2.68 m^3 s^{-1}$, $NSE = 0.75$, and $RE = 0.95$ for the River Exe at Pixton. A subset of the observed and predicted streamflow, specifically for the year 2012 when significant streamflow events occurred, is shown in Figure 5. This example demonstrates the RMSE's scaling issues, given that it is greater for the Avon at Bathford than for the Exe at Pixton due to the former's higher capacity and average streamflow, yet the overall fit for the Avon at Bathford is better, as evidenced by the better NSE score. Our metric, RE, indicates that the driver of this error for both rivers is weaker model performance around the extremes, with this being more problematic for the River Exe, quantifying that relative inability to predict the Spring and Winter extremes in the year 2012.

We have demonstrated in this paper the ability of RE to provide quantification of the shape of model error, both with unrealistic synthetic and real-world data. In the real-world context, by utilizing RE conjunction with existing error metrics, RE provides important quantification of the distribution of error relative to the observed dataset. This has the potential to enable more targeted model optimization. Furthermore, at the end of Section 2.2, we described the adaptability of RE to focus on one tail; in practice, hydrology is one area where this approach could be more suitable if investigating floods and droughts separately.

Although we did not test RE on discrete probability distributions, the extension to such situations is such that the weight function, $\Psi(y)$, generates a series of discrete weights, rather than a continuous one. However, for uniformly distributed data, the value of $\Psi(y)$ is trivially either 1 or 0 and is, consequently, only sensitive to errors outside of the bounds of the distribution.

5. Reflective loss function

We further extend this framework to the loss function in a neural network. The mean squared error (MSE) is a commonly used loss function in neural networks (Hastie et al., 2009; Bishop, 2016) to which our modification is easily applied, specifically the unnormalized form of RE. This loss function, L_{RE}, we propose for training parametric machine learning models, in place of the often used MSE, it increases the relative penalty applied to extremes and is expressed in Equation 5.1.

$$L_{RE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i'\right)^2 \cdot \Psi(y_i) \tag{5.1}$$

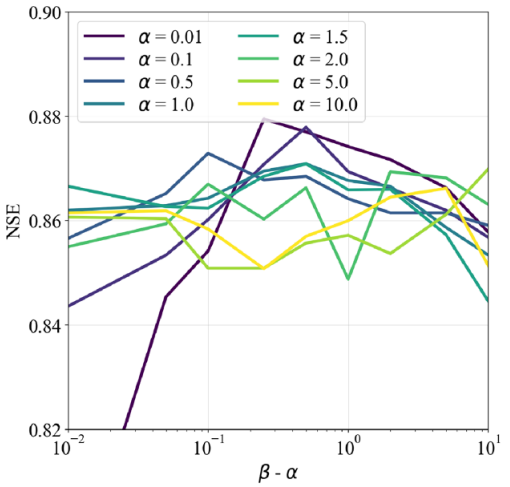
Whereas for the error metric, we let $\alpha = 1$ and $\beta = 1$, we now consider α and β as tunable hyperparameters. Thus, we can adjust the penalty applied to routine and extreme data. In order to minimize the routine loss and maximize the extreme loss, the condition $\alpha \leq \beta$ must be true.

$$\Psi(y) = -\alpha \cdot \frac{U(y)}{\kappa} + \beta \tag{5.2}$$

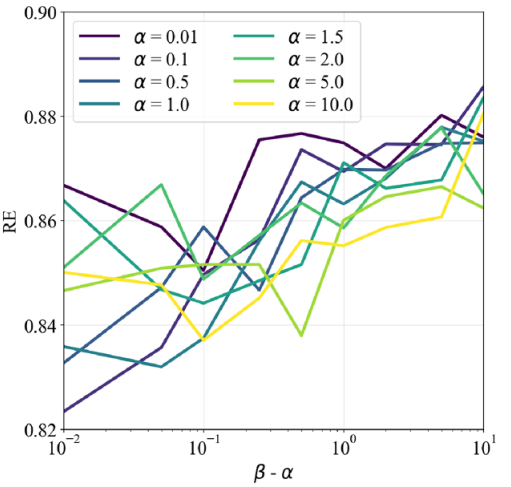
https://doi.org/10.1017/eds.2025.16 Published online by Cambridge University Press

To test the loss function, we utilize the same hydrological problem as described in the previous section, modifying the values for α and β using a grid search method whilst keeping all other aspects of network architecture and training the same, such as the training and test set compositions. The results different α and β pairs are shown in Figure 6, in terms of the difference between β and α.

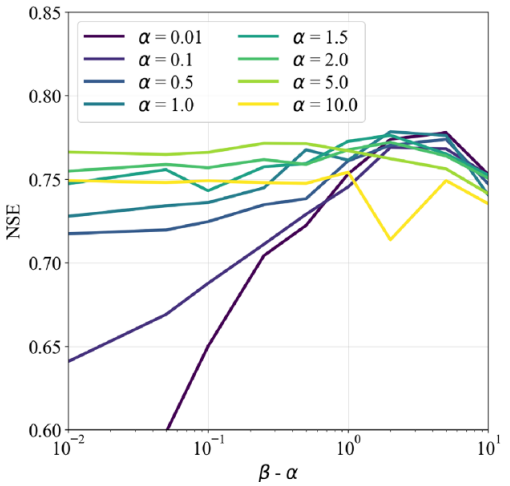
Our objective is to identify values that improve NSE whilst also reducing RE. From the investigated pairs of α and β, low values of α resulted in poor performance due to their adverse impact on the back propagation algorithm. At suitably large α, where $\alpha \geq \frac{1}{2}$, and for the parameter values where the loss function would minimize the error for routine data to 0, that is where $\alpha = \beta$, the network achieves an improvement in RE for a minor reduction in the standard error metrics. However, for the cases where



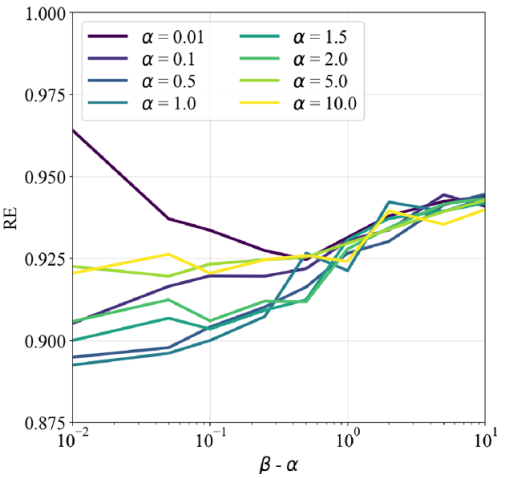
(a) NSE evaluated at different α and β pairs for model test set performance on the River Avon



(b) RE evaluated at different α and β pairs for model test set performance on the River Avon



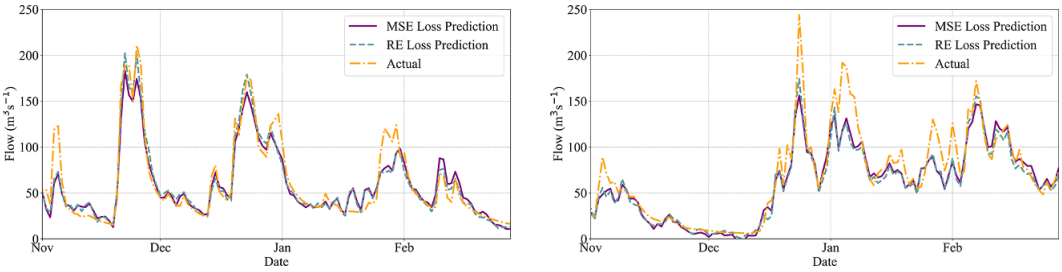
(a) NSE evaluated at different α and β pairs for model test set performance on the River Exe



(b) RE evaluated at different α and β pairs for model test set performance on the River Exe

Figure 6. *Performance in terms of NSE and RE for different α and β pairs in the Reflective Loss Function used for training a neural network on streamflow data from the Rivers Avon ((a) and (b)) & Exe ((c) and (d)).*

https://doi.org/10.1017/eds.2025.16 Published online by Cambridge University Press



(a) Observations and predictions for the River Avon, gauged at Bathford for the 2012-2013 winter period

(b) Observations and predictions for the River Avon, gauged at Bathford for the 2013-2014 winter period

Figure 7. *Focused time series for winter periods for the River Avon showing observations and predictions generated from the Neural Network model using both the MSE Loss and RE Loss functions.*

$\alpha < \beta$, we achieve an improvement in RE without sacrificing overall performance. In other words, we achieve better extreme predictive capability over that for routine data. For most values of α, the improvement between different pair values of α and β, for $\alpha < \beta$, is more marginal but these combinations achieve better performance than where $\alpha = \beta$. The best results are obtained for $0.5 \leq \alpha \leq 5$ and at $0.5 \leq (\beta - \alpha) \leq 1$. Therefore, when optimizing neural models using RE as a Loss function, values such as $\alpha = 1$ and $\beta = 2$ provide improved results. Using a hyperparameter optimization framework, such as Bayesian optimization (Snoek et al., 2012), might be a more expedient route to identifying optimal values for α and β pairs than the grid search implemented here.

We further demonstrate this by showing improved predictive performance for some of the peaks in the 2012–2013 and 2013–2014 winter periods from the RE Loss model, with parameters $\alpha = 1$ $\beta = 2$, compared to the MSE Loss model in Figure 7. Although the improved extreme performance is noticeable at extremes, we remain cognizant of the fact that this method cannot address deficiencies in the representation of extremes within the training data. For example, instances such as the peaks in the 2012 period for the River Exe at Pixton are not represented within the training data at all and addressing issues with both the spatial resolution of and bias in the precipitation data could further correct extreme performance; further analysis on the cause of these extremes is not presented within this study but it does highlight the need for domain expertise in the construction of modeling frameworks.

6. Conclusion

The RE metric we have presented here is not intended to supplant the existing error metrics; indeed, it does not directly provide quantification of the magnitude of error. Instead, our metric enables the quantification of how a model's performance is distributed, such as around the extremes of a dataset, and supplements quantification of the magnitude of error. Therefore, we recommend that RE is used alongside existing error metrics to provide a more comprehensive view of model performance.

RE can help to provide a method of quantifying relative extreme performance through a single robust and coherent number in real-world data problems, such as the hydrological problem we presented. This could extend to all manner of fields where predicting the response of systems is of significantly more value around extremes, such as the treatment of the impacts from meteorological phenomena in a nonstationary climate, where the probability distributions are subject to change; for example, there is Bhatia et al. (2019); Murakami et al. (2020); Yoshida et al. (2017).

We also presented a framework for tailoring the optimization of machine learning models, specifically neural models, to enable better prediction of extremes. Again, this provides an improvement that should not come at the expense of other facets of machine learning model training, such as ensuring representation of extremes within the data, but this is an adaptation that can be included to help drive performance toward extremes at minimal increase in complexity as far as the practitioner might be concerned.

https://doi.org/10.1017/eds.2025.16 Published online by Cambridge University Press

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2025.16.

Author contributions. Conceptualization: R.E.R. and H.M.; Data Curation: R.E.R.; Funding acquisition: R.E.R., S.H., A.M., and E.S.; Methodology: R.E.R. and H.M.; Project administration: R.E.R., S.H., A.M., and E.S.; Software: R.E.R.; Supervision: S.H., A.M., and E.S.; Visualization: R.E.R. and H.M.; Writing—Original draft: R.E.R. and H.M.; Writing—Review and editing: R.E.R. and H.M.

Competing interests. The authors declare no competing interests exist.

Data availability statement. The data required to reproduce these results are available at the National River Flow Archive Dixon et al. (2013) using the data portal https://nrfa.ceh.ac.uk/data/search or API https://nrfaapps.ceh.ac.uk/nrfa/nrfa-api.html and from the Copernicus Climate Data Store https://doi.org/10.24381/cds.143582cf. We have made the code to run the model available at https://doi.org/10.5281/zenodo.14933256.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. This research was supported by a grant from the Engineering and Physical Sciences Research Council (EP/R512461/1). This research was supported by a Fellowship from the Royal Commission for the Exhibition of 1851.

References

Abrahart RJ, **Anctil F**, **Coulibaly P**, **Dawson CW**, **Mount NJ**, **See LM**, **Shamseldin AY**, **Solomatine DP**, **Toth E and Wilby RL** (2012) Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography: Earth and Environment 36*(4),480–513. https://doi.org/10.1177/0309133312444943.

Barriopedro D, **Fischer EM**, **Luterbacher J**, **Trigo RM and García-Herrera R** (2011) The hot summer of 2010: Redrawing the temperature record map of Europe. *Science 332*(6026), 220–224. https://doi.org/10.1126/science.1201224.

Bell W (1984) Signal extraction for nonstationary time series. *The Annals of Statistics 12*(2),646–664. http://www.jstor.org/stable/2241400.

Besaw LE, **Rizzo DM**, **Bierman PR and Hackett WR** (2010) Advances in ungauged streamflow prediction using artificial neural networks. *Journal of Hydrology 386*(1–4),27–37. https://doi.org/10.1016/j.jhydrol.2010.02.037.

Bezdek JC, **Keller JM**, **Krishnapuram R**, **Kuncheva LI and Pal NR** (1999) Will the real iris data please stand up? *IEEE Transactions on Fuzzy Systems 7*(3), 368–369. https://doi.org/10.1109/91.771092.

Bhatia KT, **Vecchi GA**, **Knutson TR**, **Murakami H**, **Kossin J**, **Dixon KW and Whitlock CE** (2019) Recent increases in tropical cyclone intensification rates. *Nature Communications 10*(1), 635. https://doi.org/10.1038/s41467-019-08471-z.

Bishop CM (2016) *Pattern Recognition and Machine Learning. Information Science and Statistics*. New York: Springer New York.

Boulaguiem Y, **Zscheischler J**, **Vignotto E**, **Van Der Wiel K and Engelke S** (2022) Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks. *Environmental Data Science 1*, e5. https://doi.org/10.1017/eds.2022.4.

Box GEP and Tiao George C (1965) A change in level of a non-stationary time series. *Biometrika 52*(1/2), 181–192. https://doi.org/10.2307/2333823.

Changnon SA, **Pielke RA**, **Changnon D**, **Sylves RT and Pulwarty R** (2000) Human factors explain the increased losses from weather and climate extremes. *Bulletin of the American Meteorological Society 81*(3), 437–442 https://doi.org/10.1175/1520-0477(2000)081<0437:HFETIL>2.3.CO;2.

Chatterton J, **Clarke C**, **Daly E**, **Dawks S**, **Elding C**, **Fenn T**, **Hick E**, **Miller J**, **Morris J**, **Ogunyoye F and Salado R** (2016) *The Costs and Impacts of the Winter 2013 to 2014 Floods*. Technical Report SC140025/R1, Bristol: Environment Agency. https://assets.publishing.service.gov.uk/media/603549118fa8f5480a5386be/The_costs_and_impacts_of_the_winter_2013_to_2014_floods_-_report.pdf.

Dawson CW, **Abrahart RJ**, **Shamseldin AY and Wilby RL** (2006) Flood estimation at ungauged sites using artificial neural networks. *Journal of Hydrology 319*(1–4), 391–409. https://doi.org/10.1016/j.jhydrol.2005.07.032.

Ding D, **Zhang M**, **Pan X**, **Yang M and He X** (2019) Modeling extreme events in time series prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage, AK: ACM, pp. 1114–1122. https://doi.org/10.1145/3292500.3330896.

Dixon H, **Hannaford J and Fry MJ** (2013) The effective management of national hydrometric data: experiences from the United Kingdom. *Hydrological Sciences Journal 58*(7), 1383–1399. https://doi.org/10.1080/02626667.2013.787486.

Easterling DR (2000) Climate extremes: Observations, modeling, and impacts. *Science 289*(5487),2068–2074. https://doi.org/10.1126/science.289.5487.2068.

Faulkner D (2008) *Rainfall Frequency Estimation. Number 2 in Flood Estimation Handbook/Institute of Hydrology.* Wallingford: Centre for Ecology and Hydrology.

Fouillet A, **Rey G**, **Laurent F**, **Pavillon G**, **Bellec S**, **Guihenneuc-Jouyaux C**, **Clavel J**, **Jougla E and Hémon D** (2006) Excess mortality related to the August 2003 heat wave in France. *International Archives of Occupational and Environmental Health 80*(1), 16–24. https://doi.org/10.1007/s00420-006-0089-4.

Govindaraju RS, **Ramachandra Rao A and Singh VP** (eds.) (2000) *Artificial Neural Networks in Hydrology, volume 36 of Water Science and Technology Library*. Netherlands, Dordrecht: Springer. https://doi.org/10.1007/978-94-015-9341-0.

Gupta H V, **Kling H**, **Yilmaz KK and Martinez GF** (2009) Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology 377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

Haan L and Ferreira A (2006) *Extreme Value Theory: An Introduction. Springer Series in Operations Research*. New York; London: Springer.

Hastie T, **Tibshirani R and Friedman J** (2009) *The Elements of Statistical Learning. Springer Series in Statistics*. New York: Springer. https://doi.org/10.1007/978-0-387-84858-7.

Hersbach H, **Bell B**, **Berrisford P**, **Biavati G**, **Horányi A**, **Muñoz-Sabater J**, **Nicolas J**, **Peubey C**, **Radu R**, **Rozum I**, **Schepers D**, **Simmons A**, **Soci C**, **Dee D and Thépaut J-N** (2023) ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.adbb2d47.

Hersbach H, **Bell B**, **Berrisford P**, **Hirahara S**, **Horányi A**, **Muñoz-Sabater J**, **Nicolas J**, **Peubey C**, **Radu R**, **Schepers D**, **Simmons A**, **Soci C**, **Abdalla S**, **Abellan X**, **Balsamo G**, **Bechtold P**, **Biavati G**, **Bidlot J**, **Bonavita M**, **Chiara G**, **Dahlgren P**, **Dee D**, **Diamantakis M**, **Dragani R**, **Flemming J**, **Forbes R**, **Fuentes M**, **Geer A**, **Haimberger L**, **Healy S**, **Hogan RJ**, **Hólm E**, **Janisková M**, **Keeley S**, **Laloyaux P**, **Lopez P**, **Lupu C**, **Radnoti G**, **Rosnay P**, **Rozum I**, **Vamborg F**, **Villaume S and Thépaut J-N** (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society 146*(730), 1999–2049. https://doi.org/10.1002/qj.3803.

Jonkman SN, **Maaskant B**, **Boyd E and Levitan ML** (2009) Loss of life caused by the flooding of New Orleans after Hurricane Katrina: Analysis of the relationship between flood characteristics and mortality. *Risk Analysis 29*(5), 676–698. https://doi.org/10.1111/j.1539-6924.2008.01190.x.

Kendon M and McCarthy M (2015) The UK's wet and stormy winter of 2013/2014. *Weather 70*(2), 40–47. https://doi.org/10.1002/wea.2465.

Kingma DP and Ba J (2017) Adam: A Method for Stochastic Optimization, January. Available at http://arxiv.org/abs/1412.6980.

Muchan K, **Lewis M**, **Hannaford J and Parry S** (2015) The winter storms of 2013/2014 in the UK: hydrological responses and impacts. *Weather 70*(2), 55–61. https://doi.org/10.1002/wea.2469.

Murakami H, **Delworth TL**, **Cooke WF**, **Zhao M**, **Xiang B and Hsu P-C** (2020) Detected climatic change in global distribution of tropical cyclones. *Proceedings of the National Academy of Sciences 117*(20), 10706–10714. https://doi.org/10.1073/pnas.1922500117.

Nash JE and Sutcliffe JV (1970) River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology 10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Parry S, **Marsh T and Kendon M** (2013) 2012: From drought to floods in England and Wales. *Weather 68*(10), 268–274. https://doi.org/10.1002/wea.2152.

Proakis JG and Manolakis DG (2007) *Digital Signal Processing: Principles, Algorithms, and Applications*, 4th Edn. Upper Saddle River, NJ: Pearson/Prentice Hall.

Ramachandran P, **Zoph B and Le QV** (2017, October) Searching for Activation Functions. Available at http://arxiv.org/abs/1710.05941.

Ribeiro RP and Moniz N (2020) Imbalanced regression and extreme value prediction. *Machine Learning 109*(9–10), 1803–1835. https://doi.org/10.1007/s10994-020-05900-9.

Ritter A and Muñoz-Carpena R (2013) Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology 480*, 33–45. https://doi.org/10.1016/j.jhydrol.2012.12.004.

Rouse RE, **Khamis D**, **Hosking S**, **McRobie A and Shuckburgh E** (2025) Streamflow prediction using artificial neural networks and soil moisture proxies. *Environmental Data Science 4*, e5. https://doi.org/10.1017/eds.2024.48.

Rumelhart DE, **Hinton GE and Williams RJ** (1986) Learning representations by back-propagating errors. *Nature 323*(6088), 533–536. https://doi.org/10.1038/323533a0.

Siffer A, **Fouque P-A**, **Termier A and Largouet C** (2017) Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax, NS: ACM, pp. 1067–1075. https://doi.org/10.1145/3097983.3098144.

Sigillito V, **Wing S**, **Hutton L and Baker K** (1989) Ionosphere [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5W01B.

Snoek J, **Larochelle H and Adams RP** (2012) Practical bayesian optimization of machine learning algorithms. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2, NIPS'12*, Lake Tahoe, Nevada. Red Hook, NY: Curran Associates Inc, pp. 2951–2959.

Tarek M, **Brissette FP and Arsenault R** (2020) Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. *Hydrology and Earth System Sciences 24*(5), 2527–2544. https://doi.org/10.5194/hess-24-2527-2020.

The Environment Agency (2013) Floods cost UK economy nearly £600 million. million Environment Agency, 26 November 2013. Archived on 20 March 2014. Available at The National Archives: https://webarchive.nationalarchives.gov.uk/ukgwa/20140320171631/http://www.environment-agency.gov.uk/news/150962.aspx.

Torgo L and Ribeiro R (2007) Utility-based regression. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, ECMLPKDD'07, Warsaw, Poland. Berlin, Heidelberg: Springer-Verlag, pp. 597–604.

https://doi.org/10.1017/eds.2025.16 Published online by Cambridge University Press

UK Centre for Ecology & Hydrology (2022, June) UK National River Flow Archive Data. Available at https://nrfa.ceh.ac.uk/data; https://nrfa.ceh.ac.uk/data.

Masson-Delmotte V, **Zhai P**, **Pirani A**, **Connors SL**, **Péan C**, **Berger S**, **Caud N**, **Chen Y**, **Goldfarb L**, **Gomis MI**, **Huang M**, **Leitzell K**, **Lonnoy E**, **Matthews JBR**, **Maycock TK**, **Waterfield T**, **Yelekçi O**, **Yu R and Zhou B** (eds) 2021 *IPCC, 2021: Summary for Policymakers*. Technical Report, Cambridge, United Kingdom: Cambridge University Press. https://doi.org/10.1017/9781009157896.001.

Wu Z, **Huang NE**, **Long SR and Peng C-K** (2007) On the trend, detrending, and variability of nonlinear and nonstationary time series. *Proceedings of the National Academy of Sciences 104*(38), 14889–14894. https://doi.org/10.1073/pnas.0701020104.

Yoshida K, **Sugi M**, **Mizuta R**, **Murakami H and Ishii M** (2017) Future changes in tropical cyclone activity in high-resolution large-ensemble simulations. *Geophysical Research Letters 44*(19), 9910–9917. https://doi.org/10.1002/2017GL075058.

Cite this article: Rouse RE, Moss H, Hosking S, McRobie A and Shuckburgh E (2025). Reflective error: a metric for assessing predictive performance at extreme events. *Environmental Data Science*, 4: e26. doi:10.1017/eds.2025.16