Technical Memo

ECMWF
European Centre for Medium-Range Weather Forecasts

935

A Hydrologist's Guide to the CRPS

Karan Ruparell
Kieran M. R. Hunt
Hannah L. Cloke
Christel Prudhomme
Florian Pappenberger
Matthew Chantry

December 9, 2025

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/publications/

Contact: library@ecmwf.int

Abstract

The Continuous Ranked Probability Score (CRPS) is a vital metric in earth system forecasting, however it is still unintuitive for many users. In hydrology, we are increasingly transitioning to probabilistic and ensemble decisions making, and so informed use of metrics such as the CRPS is crucial. In this technical memorandum, we provide an overview of the CRPS. We highlight its growing importance, particularly in the context of machine learning, and then explain its relationship to various other metrics and scores commonly used in earth system forecasting, such as the Brier score and Mean Absolute Error. We then go on to apply the CRPS in the context of hydrological forecasting.

1 Plain Language Summary

While commonly used in meteorology, the Continuous Ranked Probability Score (CRPS) is often misunderstood. This is especially true in hydrology, where deterministic forecasts are still predominantly used in decision making. In this technical memo, we offer different intuitions for thinking about the CRPS, looking at both its discrete and continuous forms, and its relationship to the Brier Score. We show extensions of the CRPS, and how the CRPS can be used in a hydrological context to highlight the value of ensembles over individual daily quantiles.

2 What is this technical memo about

The goal of this technical memo is to provide some intuition on what the CRPS is, and how it relates to other metrics such as the Brier Score and the Mean Absolute Error. In doing so, we hope that users will be more comfortable using the CRPS, and in particular more comfortable using it to evaluate a greater range of forecast traits. This will be especially important as we continue to invest in Machine Learning forecasting models, where we need to test the physical realism of our models with greater scrutiny Lang et al. [2024].

We will explain its extensions in the Energy Score and Variogram Score, as well as its limitations. At the end, we will provide some toy examples in the context of hydrological forecasting.

Why is this relevant now: why the CRPS is becoming more important

In most Physics-based models, we first create physically sound simulations, and then the inputs are perturbed to generate ensemble forecasts Alfieri et al. [2013]. This means they are lead-time unaware. This has two benefits. First, the physical properties of the prediction target, such as the water balance equation for a hydrological model, are forced to be preserved. This limits the amount of testing that needs to be done on the physical realness of the models, such as on the variance of flow or total volume released in a week. Secondly, because simulations have a lead time of 0, the models can be seen as creating a single, very long, timeseries for each grid-cell. So when looking at the physical accuracy of our physical model, we can look at properties of a long timeseries, such a full hydrological year, which allows us to look at processes such as the rising limb, or flow duration curve.

This is not necessarily true for machine learning or post-processed models. Machine learning forecasts may be trained using other forecast data as inputs (Nearing et al. [2023]), or the loss function used to

train them may be lead time aware. This means they won't by default rely on previous predictions as truth, so they won't necessarily preserve things like the water balance equation between the forecast at two timesteps. In post-processing physical models, we tend to do this to minimise some error function, in a way that leads to different outcomes at different lead times, and so again we may not preserve physical properties. In either of these cases, there is no guarantee that we will produce physically plausible trajectories, where the string of predictions from time t to $t+n$ is realistic. There are many ways a timeseries of a variable can be incorrect, as the individual days are highly interdependent, so we need a way both to train and to test machine learning models to produce multi-lead-time forecasts. Rank histograms, narrowness, the CRPS (Equation 1), and skewness are all evaluation metrics used to evaluate the performance of the IFS (Hersbach et al. [2020].)

3 What is the CRPS

3.1 Discrete form of the CRPS

At a high level, the CRPS represents the average mean absolute error (MAE) between the observation y and the different ensemble forecast members $\hat{y}_j$, minus the average distance between forecast members, which acts to reduce mode-seeking behaviour and reward forecast spread. This is most obvious when the CRPS is written in its discrete form, as shown in Equation 1. When the number of ensemble members, M, is one, the first term is identical to the MAE, and the second term shrinks to zero.

$$\text{CRPS}\left(\{\hat{y}_j\}_{j=1}^{M}, y\right) = \underbrace{\frac{1}{M}\sum_{j=1}^{M}|\hat{y}_j - y|}_{\text{Average MAE}} - \underbrace{\frac{1}{2M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}|\hat{y}_j - \hat{y}_k|}_{\text{Average Spread}} \tag{1}$$

Viewing the CRPS in this way gives us our first insight in how to use it: as an ensemble equivalent of the Mean Absolute Error (MAE), and by extension the Mean Squared Error. This means we can also use and alter the CRPS in the same way we would the MAE or MSE.

It is very popular, for example, to use the MAE or MSE to test the ability of our model to predict various properties of our forecast, as in Equations 2 and 3, beyond the standard variables such as precipitation or temperature. This might be done when analysing a function of multiple variables, for example the Penman Monteith equation (Allen et al. [1998]) for evapotranspiration. Or we might look at a summary of how a variable changes over time as we do when looking at the alpha term of the Kling Gupta Efficiency (Equation12) [Gupta et al., 2009], or of some index variable. Generally, we can write these as functions $g(\cdot)$, where $\cdot$ is either an ensemble forecast $\hat{y}$ or observations y, and $g(\cdot)$ is a function that take in a forecast or observation and returns a single value. For any such function $g(\cdot)$, where y is the observed time series, we can define a metric to evaluate a deterministic forecast or ensemble mean $\hat{y}$, by calculating the mean squared error of $g(\hat{y})$. We can do the same thing with the CRPS, getting the CRPS of $g(\hat{\mathbf{y}})$, why $\hat{\mathbf{y}}$ is the ensemble of forecasts, and $g(\hat{\mathbf{y}})$ is the set of forecasts for $g(y)$ that comes from using each of these ensemble members to predict $g(y)$.

$$\text{MAE}_g(\hat{y}, y) = |g(\hat{y}) - g(y)| \tag{2}$$

$$\text{MSE}_g(\hat{y}, y) = (g(\hat{y}) - g(y))^2 \tag{3}$$

$$\text{CRPS}_\text{g}(\hat{\mathbf{y}}, y) = \text{CRPS}(g(\hat{\mathbf{y}}), g(y)) \tag{4}$$

Going further, we might standardise our metrics, by dividing the score by some reference forecast $\hat{y}_{ref}$. For the MSE this can be seen in Equation 5 . For any function $g(\cdot)$, and a reference ensemble forecasts $\hat{\mathbf{y}}_{ref}$ we can similarly define the $CRPS_g$, and a skill score for it, the Continuous Ranked Probability Skill Score (CRPSS), or $CRPSS_g$ shown in Equation 6

$$\text{Skill-MSE}_g(\hat{y}, y) = 1 - \frac{(g(\hat{y}) - g(y))^2}{[g(\hat{y}) - g(\hat{\text{y}}_\text{ref})]^2} \tag{5}$$

$$\text{CRPSS}_\text{g}(\hat{\mathbf{y}}, \hat{\mathbf{y}}_{ref}, y) = 1 - \frac{\text{CRPS}_g(\hat{\mathbf{y}}, y)}{\text{CRPS}_g(\hat{\mathbf{y}}_{ref}, y)} \tag{6}$$

Being able to manipulate the MSE (Equation 3) in this way forms the basis of each component of the Kling-Gupta Efficiency (KGE) (Equation 12) in hydrological forecasting, as well as for the Nash-Sutcliffe Efficiency Nash and Sutcliffe [1970]. The R^2 and $RMSE$ similarly rely on the MSE. We can use the same intuitions we have for the MAE and MSE to use the CRPS in more beneficial ways.

Looking at the discrete form alone, it might not be clear that the CRPS is minimised only by a model that perfectly captures the true distribution of the target variable. Such metrics are often called 'proper scoring rules'. Since we are taking away the ensemble spread in the error function, what happens if we make the ensemble spread incredibly wide? Would that lead to a very low loss? It turns out that the CRPS is in fact a proper scoring rule Pic et al. [2025], and so we can safely use it to evaluate our models. To best see this, we can look at the continuous form of the CRPS.

3.2 Continuous form of the CRPS

Another way of looking at the CRPS is as the area between two cumulative distribution functions (CDF), as in Figure 1. In particular, it is the area between the empirical CDF produced by the ensemble forecast, F ,and the indicator function $g(x) = 0$ if $x < y$ and $\mathbf{1}(x) = 1$ otherwise, (Equation 7). The equivalence between this form and the discrete form above can be found in Baringhaus and Franz [2004]. The integral form shows us that we can think of the CRPS in a similar way to other metrics that compare two distributions, such as the Kolmogorov-Smirnov (KS) test. The key difference is that the CRPS sums the differences between the two cumulative distributions at all points, whereas the KS test only looks at the point where the difference between the two distributions is greatest. The Kullback-Leibler divergence is a common metric in machine learning that is similar to the CRPS, but is less intuitive.

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbf{1}\{z \geq y\})^2 \, dz \tag{7}$$

If we think about evaluating the inside of the integral of the CRPS at a single point, z_0 we reach the Brier Score (BS), as shown in Equation 8. When we have $C - 1$ thresholds, we get the Ranked Probability Score (RPS) in Equation 9. If we keep adding thresholds, eventually having one at every possible value of z_0, the summation will become an integral, and be equal to the CRPS. Another intuition for the CRPS, then, is that it is the continuous form of the Brier Score. For those familiar with how to use the Brier

Score, the CRPS can be seen as the equivalent, when we wish to know more about the overall distribution as opposed to a particular threshold. This difference between the Brier Score and the CRPS also reveals

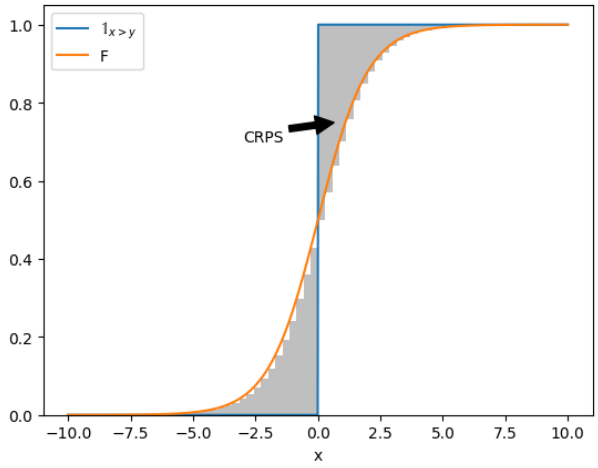


Figure 1: Diagram showing the CRPS as the area between the observed CDF of observations (an indicator function in the case of a single target) and the empirical CDF generated by an ensemble

the major strength and weakness of the CRPS: the CRPS is a great summary of the average performance of an ensemble, but it doesn't tell you about any particular threshold or physical property.

$$\mathrm{BS}(F,y) = (F(z_0) - \mathbf{1}\{y \geq z_0\})^2 \tag{8}$$

$$\mathrm{RPS}(F,y) = \sum_{k=1}^{C-1} (F(z_k) - \mathbf{1}\{y \geq z_k\})^2 \tag{9}$$

4 The CRPS is great, the CRPS isn't perfect

As we saw previously, the CRPS provides a single value metric to evaluate the overall performance of an ensemble, and is a very intuitive way of defining such a metric. However, by nature of being a single value that combines different properties, it cannot be easily used to identify the specific cause of ensemble error. A high CRPS could be a result of any number of factors.

It could mean that the ensemble variance is too high, or that it is too low. It could mean that that there is a systematic bias in the forecast, or that the average ensemble member is just not very skilled without there being a systematic bias, or it could mean that the ensemble occasionally has a member that is exceptionally wrong, increasing the error dramatically. Just like the MSE, or the Nash-Sutcliffe Efficiency (NSE), or the King-Gupta Efficiency (KGE), the CRPS is not a total solution, and other metrics are still very important. To test average performance, we might get the MSE of the ensemble mean, or we could look at the bias of the ensemble mean. To look at the spread, we may use narrowness, or we may judge our ensemble visually using a rank histogram. If we want to evaluate our models against forecast thresholds, we can use the Brier Score. If we only want to look at extreme ensemble members, we can use the quantile loss. The CRPS is certainly not the only way for forecasters to evaluate their forecast

models, but it is an important way. With it, we can quickly communicate the general performance of an ensemble, and then look deeper.

5 Extensions to the CRPS

The CRPS is a score that calculates the error for a single prediction variable, or transformed variable. However, in most of our use cases there are multiple variables we wish to predict. This may be different types of variables, such as river flow and precipitation, or it might be the same variable over different lead times or areas. When we make ensemble predictions for multiple variables at once, we would like each forecast to be a realistic combination of these variables, which is stricter than each variable having realistic ensemble timeseries on its own.

Specifically, consider what the river flow might be at the end of a large catchment, over the course of two days. If it rains heavily, the catchment might have quite high flow on both days. If it doesn't rain at all, then we might expect the river flow to be relatively low on both days. It is very unlikely, however, for the river flow to be very high one day and low the next, even if each of those individual outcomes is fairly likely. In all of our forecasts, we are really forecasting over the joint probability of many observations over space and time, as opposed to predicting many independent values.

The Energy Score (ES) is the multivariate form of the CRPS, and it looks at the expected distance between two sets of observations. Just like the CRPS looks at the mean absolute error between a forecast and an observation of the same variable, the Energy Score looks at the Euclidean distance between the different forecast variables, and compares that to the spread of the forecast members. The formula for the Energy Score is shown in Equation 10. Intuitively, the Energy Score calculates the expected distance of a forecast member $\hat{Y}$ to the observation Y, and compares that to the expected pairwise distance between forecast members $\hat{Y}$ and $\hat{Y}'$. For a perfect forecast, the expected difference between a forecast member and the observation is equivalent to half the distance between two sampled forecast members. In this case, the observation is indistinguishable from the ensemble. The Energy Score also has a β term, which is used to control how biased the score is towards extreme errors. The Energy Score allows us to consider the difference between two multivariate forecasts, however, it requires much more data than the CRPS to be useful. This is because having more variables to consider increases the dimensionality of the problem, which means our score can be more noisy.

$$ES_\beta(D,Y) = \underbrace{\mathbb{E}_{\hat{Y}\sim D}\left[\|\hat{Y}-Y\|_2^\beta\right]}_{\text{Average distance to observation}} - \frac{1}{2}\underbrace{\mathbb{E}_{\hat{Y},\hat{Y}'\sim D}\left[\|\hat{Y}-\hat{Y}'\|_2^\beta\right]}_{\text{Average spread}} \tag{10}$$

Instead, the Variogram Score (VS), shown in Equation 11, is often used Scheuerer and Hamill [2015] in meteorological forecasting. Instead of looking at the error of a forecast, it looks at the difference between different variables i and j, and compares the difference between the observations Y_i, Y_j with the average difference between the forecasted values $\hat{Y}_i, \hat{Y}_j$. This is different from the Energy Score, as it evaluate how good the model is at predicting the spread of the different variables, whereas the Energy Score evaluates how good the model is at predicting each of the variables themselves. The $w_{i,j}$ term allows us to weight the difference between certain variables more highly, for example if we know that they should be highly correlated. The p term allows us to control for how much we weight larger errors compared to smaller errors, with common choices being $p = 0.5$ or $p = 1$. In order to compare different parameters i and j with each other, the variables often need to be standardised.

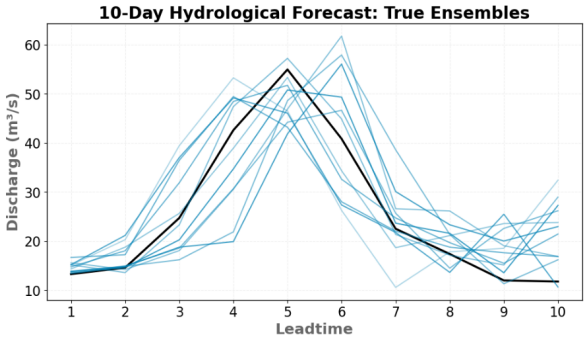
$$VS_p(D,Y) = \sum_{i,j=1}^{n} w_{ij} \left(\underbrace{|Y_i - Y_j|^p}_{\text{Observed difference}} - \underbrace{\mathbb{E}_{\hat{Y}\sim D}\left[|\hat{Y}_i - \hat{Y}_j|^p\right]}_{\text{Predicted difference}} \right)^2 \tag{11}$$

6 How can the CRPS apply to Hydrological Forecasting?

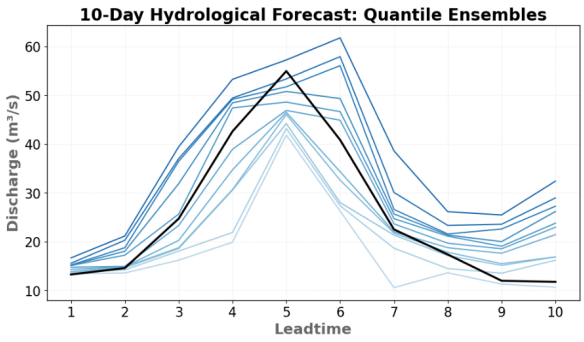
With our new insights on how to use the CRPS, how might this apply to hydrological forecasting?

6.1 Example Forecasts

To show why evaluating the temporal relationship of our ensembles matter, consider an artificially generated 10-day hydrograph with an 11-member ensemble, shown in Figure 2. On the left is the true ensemble forecast, where each member represents a complete trajectory. On the right is what the forecast looks like if we construct *pseudo-ensembles* using the quantiles at each timestep. Looking at the true ensemble, we see that the model has uncertainty in the timing of flow - members predict a peak on different days, but with similar total volumes over the course of the ten days. If we used the quantile representation as if they were ensembles, which is often done intuitively, we see something completely different. The quantiles show massive uncertainty in whether we get high or low flow overall, whereas there seems to be much greater consensus on when the peak day will be.



(a) True ensemble trajectories



(b) Quantile-based pseudo-ensembles

Figure 2: Simulated comparison of an 11-member ensemble hydrograph forecast, shown in blue, and the observed flow, shown in black. **(a)** The true ensemble members, showing complete trajectories. Note the primary uncertainty is in the timing of the peak flow. **(b)** Pseudo-ensemble trajectories constructed from the quantiles at each timestep. This common but misleading representation incorrectly implies uncertainty in the magnitude of the flow, while showing a false consensus on the peak timing.

Since each forecast and quantile-based members share the same predictions each timestep, the per-day CRPS is the same. Both score between 15-45 m³/s depending on the day, with the highest uncertainty around the peak. The per-day CRPS correctly tells us which days are hardest to predict, but it misses fundamental differences in the two forecasts.

The Kling-Gupta Efficiency (Equation 12) is a metric that captures some of these fundamental differences. Just like the CRPS, it does not identify a single form of model error, but rather three different forms. It is a skill score, that considers how accurately the model predicts the mean flow over a hydro-

graph, β, how accurately the model predicts the variance of the hydrograph, α, and how accurately the model predicts the timing of events, i.e. the correlation between the predicted and observed hydrograph, r. In order to study each of these traits independently, we can compute the CRPS of the α and β. For a general hydrological signature p, we can compute $CRPS_p$ as in Equation 13.

$$KGE = 1 - \sqrt{\underbrace{(r-1)^2}_{\text{Correlation}} + \underbrace{(\alpha-1)^2}_{\text{Variability}} + \underbrace{(\beta-1)^2}_{\text{Bias}}} \tag{12}$$

$$\text{CRPS}_{\text{g}}\left(\{\hat{y}_j\}_{j=1}^{M}, y\right) = \underbrace{\frac{1}{M}\sum_{j=1}^{M}\left|g(\hat{y}_j) - g(y)\right|}_{\text{Average MAE}} - \underbrace{\frac{1}{2M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}\left|g(\hat{y}_j) - g(\hat{y}_k)\right|}_{\text{Average Spread}} \tag{13}$$

Table 1: Hydrograph properties evaluated using property-based scoring. For each property $g(\cdot)$, we compute: $\text{CRPS}_g(\hat{\mathbf{y}}, y) = \text{CRPS}(g(\hat{\mathbf{y}}), g(y))$, where y is the observed hydrograph, $\hat{y}$ is the point forecast, and $\hat{\mathbf{y}}$ is the ensemble of forecasts. The Variogram Score is computed directly on the multivariate distribution and does not follow the property-based framework.

Property	**Definition** $g(y)$	**Purpose of CRPS**
Mean Flow	$g(y) = \frac{1}{T}\sum_{t=1}^{T} y_t$	Assesses accuracy in predicting the mean volume of water over the forecast period
Variance of Flow	$g(y) = \text{Var}(y)$	Captures the variability of the flow hydrograph

When we look at the CRPS of summary statistics, such as total flow or variance of flow, the differences can be seen more clearly. Table 2 shows the results. For total flow over the 10 days, the true ensemble scores 127 m³/s while the quantile pseudo-ensemble scores 438 m³/s - more than three times worse. The true ensemble members all predict similar total volumes even though they disagree on timing, while the quantile interpretation has a huge range of possible volumes. In the worst case, these combinations would not be physically possible, for example if we knew an upstream flood event had just occurred, we would expect to have more confidence in the overall flow, but less in the timing. A similar story can be seen for the variance of flow

The per-timestep CRPS alone isn't enough. By look at the CRPS of other statistics, we can identify differences between different ensemble models, and then can further inspect these differences with other tools such as narrowness or rank histograms.

CRPS Statistic	**True Ensemble**	**Quantile Pseudo-Ensemble**
Total Flow (m^3/s)	127	438
Mean Flow (m^3/s)	12.7	43.8
Variance of Flow $((m^3/s)^2)$	8.2	31.5

Table 2: CRPS for aggregate statistics. Despite identical per-timestep CRPS, the multivariate metrics show the quantile interpretation is much worse.

7 Conclusion

The Continuous Ranked Probability Score is a powerful and versatile tool for evaluating ensemble forecasts. It is also an intuitive tool, with parallels in metrics we already use frequently in hydrological forecasting, like the mean squared error, mean absolute error, brier skill score, or Kolmogorov-Smirnov test. The CRPS can be understood through all of these lenses: as an ensemble-aware MAE that penalises unrealistic spread, as the integral of Brier Scores across all thresholds, and as the a metric of the similarity between two cumulative distribution functions.

This versatility makes the CRPS particularly valuable as we continue to develop forecasting models that use machine learning. Unlike physics-based models that inherently preserve physical constraints, ML models require more rigorous testing to ensure they produce physically plausible trajectories across lead times. The CRPS provides a framework for evaluating not just individual forecasts, but the joint probability distributions that characterize multi-variable, multi-temporal predictions. While not shown here in the interest of clarity, its extensions in the Energy Score and Variogram Score may also be useful to forecasters, when we wish to ensure that our forecasts are correctly predicting multiple variables at once, and are already used in meteorological forecasting Scheuerer and Hamill [2015].

We've also shown how we can apply the CRPS to create similar metrics to familiar hydrological metrics like the NSE and KGE, demonstrating that the CRPS is not a replacement for existing evaluation frameworks, but a natural extension of them into probabilistic evaluation. But we would never rely solely on the NSE or KGE to diagnose all model deficiencies, and the CRPS too should be used alongside other diagnostic tools—rank histograms, bias metrics, and threshold-specific Brier Scores—to build a complete picture of ensemble performance.

As ensemble forecasting becomes increasingly central to our operations, and as we evaluate more complex forecast properties beyond standard variables, informed use of the CRPS will be essential. By understanding it as both an extension of familiar deterministic metrics and as a proper probabilistic scoring rule, forecasters can confidently apply it and manipulate it to evaluate the physicality, skill, and reliability of next-generation forecasting systems.

8 Acknowledgements

We would like to acknowledge Martin Leutbecher and Zied Ben Bouallegue for their contribution to this work. They were able to point out much of the literature around this topic, as well as some of the framing for this technical memorandum. This work was supported by the Advanced Frontiers for Earth System Prediction Doctoral Training Programme, funded by the University of Reading.

References

Lorenzo Alfieri, Peter Burek, Emanuel Dutra, Blazej Krzeminski, David Muraro, Jutta Thielen, and Florian Pappenberger. Glofas–global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences*, 17(3):1161–1175, 2013.

Richard G Allen, Luis S Pereira, Dirk Raes, Martin Smith, et al. Crop evapotranspiration-guidelines for computing crop water requirements-fao irrigation and drainage paper 56. *Fao, Rome*, 300(9):D05109, 1998.

Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of multivariate analysis*, 88(1):190–206, 2004.

Hoshin V Gupta, Harald Kling, Koray K Yilmaz, and Guillermo F Martinez. Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of hydrology*, 377(1-2):80–91, 2009.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Simon Lang, Mihai Alexe, Mariana CA Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D Dueben, Sara Hahner, et al. Aifs-crps: Ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *arXiv preprint arXiv:2412.15832*, 2024.

J Eamonn Nash and Jonh V Sutcliffe. River flow forecasting through conceptual models part i—a discussion of principles. *Journal of hydrology*, 10(3):282–290, 1970.

Grey Nearing, Deborah Cohen, Vusumuzi Dube, Martin Gauch, Oren Gilon, Shaun Harrigan, Avinatan Hassidim, Frederik Kratzert, Asher Metzger, Sella Nevo, Florian Pappenberger, Christel Prudhomme, Guy Shalev, Shlomo Shenzis, Tadele Tekalign, Dana Weitzner, and Matias Yoss. Ai increases global access to reliable flood forecasts. *arXiv preprint arXiv:2307.16104*, 2023.

Romain Pic, Clément Dombry, Philippe Naveau, and Maxime Taillardat. Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation. *Advances in Statistical Climatology, Meteorology and Oceanography*, 11(1):23–58, 2025.

Michael Scheuerer and Thomas M Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015.