Evaluating extreme precipitation forecasts: A threshold-weighted, spatial verification approach for comparing an AI weather prediction model against a high-resolution NWP model

Nicholas Loveday, Bureau of Meteorology
nicholas.loveday@bom.gov.au

Primary host: Tracy Hertneky, NSF NCAR and DTC
Secondary host: Molly Smith, NOAA GSL and DTC

October 14, 2025

Summary

Aim

The goal of this visit was to develop a verification approach that combines spatial methods with proper scoring rules to evaluate the performance of predictions of extremes. The goal was to demonstrate this approach by comparing an artificial intelligence weather prediction (data-driven) model against a high-resolution NWP model.

Outcomes

Several outcomes resulted from this visit:

1. A new verification method that merges two existing approaches was demonstrated. This method assesses model performance in a way that reflects how operational meteorologists may use a model. The approach has several favourable properties, which are discussed in this report.
2. We demonstrated this approach by evaluating how GraphCast-GFS, a pure AI model, performs against HRRR in predicting extreme 6-hour precipitation accumulations.
3. We made open-source Python code available for DTC staff and others to use in calculating threshold-weighted CRPS in the *scores* Python package.
4. We strengthened relationships between the Bureau of Meteorology, DTC, NSF NCAR, and NOAA GSL. We also proposed a path forward for incorporating the open source *scores* Python package, initially developed by the Bureau, into METplus. This would provide a direct pathway for Bureau scientists to contribute new methods to METplus.
5. This report will be converted into a journal article and submitted to a peer-reviewed journal.

1 Introduction

Over recent years, there has been rapid progress in the development of artificial intelligence weather prediction (AIWP) models, sometimes referred to as data-driven models (Ben Bouallègue et al., 2024).

These models are typically based on neural networks trained on reanalysis datasets. This approach contrasts with traditional numerical weather prediction (NWP) models, which rely on equations that model physical processes. One major advantage of AIWP models is that they are significantly more computationally efficient than NWP models for making predictions once they have been trained.

Most early global AIWP models are trained on ERA5 (Hersbach et al., 2020) data, and therefore have a similar or coarser grid resolution of approximately 0.25°. These include deterministic models (e.g., Keisler, 2022; Lam et al., 2023; Bi et al., 2023; Lang et al., 2024a) and ensemble models (e.g., Price et al., 2024; Lang et al., 2024b). In general, these models operate at much coarser resolutions than the high-resolution physical NWP models available to many international weather forecasting centers. However, higher-resolution, limited-area AIWP models have started to emerge more recently (e.g., Nipen et al., 2024; Adamov et al., 2025; Abdi et al., 2025), and this trend is expected to continue.

As modelling centers have access to high-resolution NWP models and with the accelerating development of higher-resolution AIWP models, there is a strong need for appropriate methodologies to compare the performance of models with differing spatial resolutions. Such comparisons are important to help meteorological agencies make informed decisions regarding future modelling strategies and to assist operational meteorologists and forecast system developers in understanding each model's strengths and weaknesses.

This raises several key questions we aim to address.

First, most benchmarking efforts of global AIWP models to date have relied on point-to-point verification methods (e.g., Rasp et al., 2024), with perhaps the exception of Radford et al. (2025a), who applied an object-based verification approach. Point-to-point methods may be appropriate for evaluating forecasts at individual locations (e.g., for end-users accessing weather forecasts via mobile apps). However, meteorologists and forecast systems often use forecasts spatially, where spatial coherence and physical realism are important.

Point-to-point verification methods tend to favour smoother forecasts than are typically observed in gridded observations (Subich et al., 2025). When spatial structure is important, point-to-point metrics can suffer from the "double penalty" effect (Ebert, 2008). This can occur when a forecast feature matches the observed shape and intensity but is slightly displaced in location, leading to it being penalised twice. It is penalised once for the false prediction at the forecast location, and once for the missed event at the observed location. This issue is particularly problematic for high-resolution models, which aim to represent fine-scale features. Lower-resolution models that fail to capture the feature at all may score better under point-to-point metrics, because they incur only a single penalty. Thus, while high-resolution NWP models may produce forecasts that appear more realistic to meteorologists, they may not perform better under standard point-to-point verification metrics (Mass et al., 2002).

To address the limitations of point-based verification, a wide range of spatial verification methods have been developed over the past two and a half decades (Gilleland et al., 2009; Dorninger et al., 2018). These methods are typically grouped into five categories: neighbourhood methods, scale-separation methods, feature-based (or object-based) methods, field-deformation methods, and distance measures. In this paper, we focus on an approach within the neighbourhood methods category (Ebert, 2009; Schwartz and Sobash, 2017). Neighbourhood methods are verification approaches that evaluate gridded forecasts that are located within a local neighbourhood of the observations.

Second, there is a pressing need to assess how well both AIWP and high-resolution NWP models predict extreme weather events. Much of the existing literature has focused on a limited number of case studies (e.g., Charlton-Perez et al., 2024; Morisseau et al., 2025). While case studies are valuable, they may be prone to selection bias or may not generalise across more cases. Focusing solely on cases where extremes occur can lead to the "forecaster's dilemma" (Lerch et al., 2017). The forecaster's dilemma arises when forecasters or modellers must choose between issuing honest forecasts and adjusting their forecasts to optimise a particular scoring rule (we refer to the latter as "hedging"). Proper scores discourage hedging and avoid the "forecaster's dilemma". A scoring rule is proper if the expected score is optimised by producing a forecast that corresponds to the forecaster's true belief (Winkler and Murphy, 1968; Gneiting and Raftery, 2007).

There are some recent examples of broader evaluation of extremes across many events. For example, Lam et al. (2023) and Ben Bouallègue et al. (2024) evaluated the discrimination ability of AIWP models to differentiate between climatological extreme temperature events and non-events. Olivetti and Messori (2024) assessed performance in forecasting temperature and wind extremes primarily by conditioning on observed extremes, which is also susceptible to the forecaster's dilemma. However, they also included an evaluation using a threshold-weighted squared error (Taggart, 2021) in their Appendix B, which avoids this problem and provides a more robust metric for evaluating extreme events. Zhang et al. (2025) evaluated several AIWP models and ECMWF's HRES model against observations that are more extreme than the observations in the 1979-2017 training period of the AIWP models. In an attempt to avoid the 'forecaster's dilemma', the authors additionally use an alternative approach to calculate a conditioned Root Mean Square Error (RMSE) which also conditions the data on forecast extremes; that is, they selected only the points where either the forecast or the observation was considered extreme and calculated RMSE on them. However, Taggart (2021) demonstrated that conditioning on both forecasts and observations does not, in fact, avoid the 'forecaster's dilemma'. These efforts have been limited to point-to-point verification methods, and there is a need for more comprehensive verification methods that assess the performance of models predicting extremes in a user-oriented spatial framework.

In this paper we demonstrate a verification approach that merges two existing methods and assesses the model in one particular framework that aligns with a specific way that operational meteorologists may use a model. The approach has several strengths including:

1. It is a spatial verification method that aligns with a particular way that a model may be used.
2. It uses proper scoring rules within a user-oriented framework.
3. Threshold-weighting of proper scoring rules allows for the evaluation of extremes and other important decision thresholds.
4. When comparing models with different spatial resolutions, no re-gridding is required.
5. Models can be evaluated against station-based observations which may be desirable in some cases.
6. It can be extended to measure discrimination ability.

Pagano et al. (2024) raised the research question, "Can spatial verification methods be extended to emphasize predictive performance for extremes without creating forecaster's dilemmas...?" The approach presented here provides one solution to this research question, enabling the evaluation of model performance across different grid resolutions for predicting extremes, within a user-oriented spatial framework that uses proper scoring rules. We evaluate the performance of an AIWP model against a high-resolution physical NWP model across 32 months of data.

2 Data

2.1 Observations

Station data were used to assess the performance of the forecast models using a spatial verification approach. Most evaluation of AIWP models has been done against ERA5, which has some biases in its precipitation fields, with the United States having a dry bias (Lavers et al., 2022). Additionally, most global AIWP models are trained against ERA5, and biases in ERA5 may propagate into these models and may go undetected if evaluated against ERA5. Some countries have alternative gridded rainfall datasets, which may serve as suitable substitutes, though these are not available everywhere. For this reason, some countries may place greater trust in verification using weather station data.

While most past efforts have evaluated these models against gridded analysis, WeatherReal (Jin et al., 2024) was recently developed to evaluate AIWP models against in situ observations. One limitation of this dataset is that it is only available for 2023, and our aim was to evaluate model performance over a longer period to assess extreme event prediction. Instead, one-minute Automated Surface Observing System (ASOS) (NWS, 1998) precipitation data were retrieved for the contiguous United States (CONUS). A map of ASOS weather station data is shown in Fig. 1. From this data, six-hour precipitation accumulations were derived. As a basic quality control measure, data were removed when one-minute accumulations exceeded 38 mm or six-hour accumulations exceeded 840 mm, which correspond to world record values (WMO, 1994). Additionally, six-hour accumulations were treated as missing when fewer than five hours of valid data existed within the six-hour period. ASOS data have previously been used to evaluate the HRRR model (e.g., Ikeda et al., 2013; Fovell and Gallagher, 2022), but, to our knowledge, have not been used to evaluate an AIWP model.

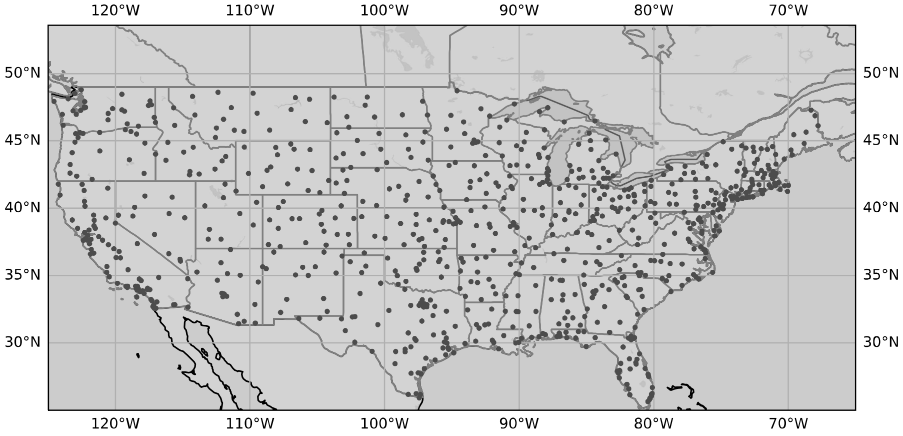

Figure 1: A map of ASOS station locations used in this study.

We also generated annual 99th and 99.9th percentile climatological thresholds for six-hour precipitation accumulations at each station to define extreme precipitation events. Since long time series are not available for all stations in the ASOS dataset, we used the grid-point that station was located in from the ERA5 reanalysis data from 1990-2020 to generate these thresholds. These thresholds will differ to those derived directly from the station data, but will be used in this framework as a reference to define extreme events.

2.2 Forecasts

We used the 00 UTC run of two different models, with base-run times spanning from 1 January 2022 to 30 August 2024 (973 model runs).

The first model we evaluated was GraphCast, with initial conditions from NOAA's Global Forecast System (GFS), using the reforecast archive generated by Radford et al. (2025b), hereafter referred to as GraphCast-GFS. The grid resolution of GraphCast-GFS is 0.25°. Radford et al. (2025a) found that, while GraphCast-GFS tended to overestimate lower rainfall amounts compared to GraphCast initialised with ECMWF's Integrated Forecasting System (GraphCast-IFS), its performance for higher rainfall amounts was similar. GraphCast-GFS was selected because it was the only AIWP model output available at the time of analysis that produced six-hour precipitation accumulations over a sufficiently long period to assess performance in predicting extremes.

Our evaluation begins in 2022 because the operational version of GraphCast was fine-tuned on data through 2021, and we sought to avoid testing the model on data used in its development. Since AIWP models can produce negative precipitation values, all negative values were set to zero, as this is a trivial post-processing step.

To compare an AIWP model with a high-resolution physical numerical weather prediction (NWP) model, we also evaluated version 4 of the High-Resolution Rapid Refresh (HRRR) model (Dowell et al., 2022). HRRR is a convection-allowing, cloud-resolving physical NWP model with a horizontal grid spacing of 3 km, which runs hourly. HRRR was selected due to its widespread use in the United States as a high-resolution operational model.

Prior to score calculation, missing data were matched between the two models to ensure a fair comparison. Since HRRR produces forecasts only up to 48 hours from the 00 UTC run, we limited the GraphCast-GFS evaluation to its first 48 forecast hours as well.

3 Using HiRA and twCRPS

This work unifies two different verification methods. We give an overview of each method before explaining how threshold weighted scores can be used in a spatial verification framework.

3.1 HiRA

One neighbourhood verification method that can be used with point observations is the High-Resolution Assessment (HiRA) framework (Mittermaier, 2014). One of the motivations for the development of the HiRA framework was to evaluate models of varying resolutions against site-based observations while avoiding the double penalty effect within a specified distance.

The HiRA framework can be summarised as follows: for a given point observation, several squares (or sometimes circles) of varying sizes are used to define neighbourhoods of forecast grid cells surrounding the observation site. Each neighbourhood can then be used to generate a pseudo-ensemble, where each grid point within the neighbourhood is considered to have an equal probability of occurring. The pseudo-ensemble can subsequently be evaluated using a probabilistic score, such as the Brier score (Brier, 1950), the ranked probability score (RPS) (Epstein, 1969), or the continuous ranked probability score (CRPS) (Matheson and Winkler, 1976). HiRA enables models with different spatial resolutions to be compared over equivalent spatial areas by selecting neighbourhoods of similar physical size. For example, if Model A has a grid resolution of 3 km and Model B has a resolution of 9 km, then a 27×27 km neighbourhood would correspond to a 9×9 grid of Model A and a 3×3 grid of Model B, both centered on the observation point. Figure 4 in Crocker et al. (2020) provides a graphical illustration of how pseudo-ensembles are constructed from neighbourhoods of various sizes.

3.2 Threshold Weighted Continuous Ranked Probability Score

In the context of evaluating the performance of forecasts for extreme events, threshold-weighted proper scoring rules are particularly important for forecast system developers and forecasters, as they help to avoid the “forecaster’s dilemma". A scoring rule is considered proper if the expected score is optimised when the forecaster issues a probability forecast that corresponds to their true belief. The continuous ranked probability score (CRPS) is a strictly proper scoring rule that can be expressed as the integral of the Brier score over all possible thresholds (Matheson and Winkler, 1976; Gneiting and Raftery, 2007).

The CRPS for evaluating a cumulative distribution function F is defined as

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(z) - \mathbb{1}\{y \leq z\}\right)^2 \, \mathrm{d}z \tag{1}$$

$$= \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F|X - X'|. \tag{2}$$

In the first expression, y is the observation, z is the decision threshold, and $\mathbb{1}\{y \leq z\}$ is the indicator function, which equals 1 if $y \leq z$ and 0 otherwise. In the second expression, X and X' are independent random variables with distribution F, and $\mathbb{E}_F$ is expectation with respect to the probability distribution F. Lemma 2.2 in Baringhaus and Franz (2004) and equation 17 in Székely and Rizzo (2005) show that the two expressions above are equal. When used to evaluate an ensemble forecast F_{ens}, the CRPS can also be expressed as

$$\mathrm{CRPS}(F_{\mathrm{ens}}, y) = \frac{1}{M}\sum_{m=1}^{M} |x_m - y| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{j=1}^{M} |x_m - x_j|, \tag{3}$$

where M is the number of ensemble members, x_m and x_j are the forecasts of a single ensemble member. Ferro (2013) developed the idea of fair scores for evaluating ensembles that provide an unbiased estimator of the CRPS as if the ensemble size were to approach infinity. This notion of fair scores is important for HiRA since the pseudo-ensembles have a different number of members depending on the neighbourhood size. The fair CRPS for evaluating an ensemble is defined as

$$\mathrm{CRPS}(F_{\mathrm{ens}}, y) = \frac{1}{M}\sum_{m=1}^{M} |x_m - y| - \frac{1}{2M(M-1)}\sum_{m=1}^{M}\sum_{j=1}^{M} |x_m - x_j|. \tag{4}$$

Weighted scoring rules have been developed to evaluate forecasts with an emphasis on specific ranges of decision thresholds, such as those that lie in the extremes of a climatological distribution. These scoring rules have proven useful in several contexts for assessing the ability to predict extreme events. For example, Loveday et al. (2024) demonstrated that the standard mean squared error (MSE) showed no statistically significant difference in temperature forecast skill between meteorologists and automated guidance. However, using a threshold-weighted MSE revealed that meteorologists performed better at predicting temperature extremes than automated guidance. Similarly, Wessel et al. (2025) showed that threshold-weighted scores can be used as a loss function to improve the predictive performance for extremes. Allen et al. (2023) illustrated how multivariate threshold-weighted scores can be applied to evaluate forecasts of rainfall accumulation over consecutive days.

Among the wide variety of threshold-weighted scores, the threshold-weighted continuous ranked probability score (twCRPS) has been the most extensively used and is defined as

$$\mathrm{twCRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(z) - \mathbb{1}\{y \leq z\}\right)^2 w(z) \, \mathrm{d}z, \tag{5}$$

where $w(z)$ is a non-negative weight function (Gneiting and Ranjan, 2011). The choice of $w(z)$ depends on the user's decision thresholds of interest. For example, if the aim is to evaluate a model's ability to produce accurate forecasts for users whose decision thresholds exceed 40°C, a suitable threshold weight function could be $w(z) = \mathbb{1}(z > 40)$. Another example might involve choosing $w(z)$ to focus on temperature ranges associated with aircraft icing, in which case the weights could vary smoothly to reflect the relative importance of different thresholds z.

Figure 2 provides a graphical illustration of how the twCRPS is computed when a constant weight is applied to the upper tail of decision thresholds, aiding interpretation.

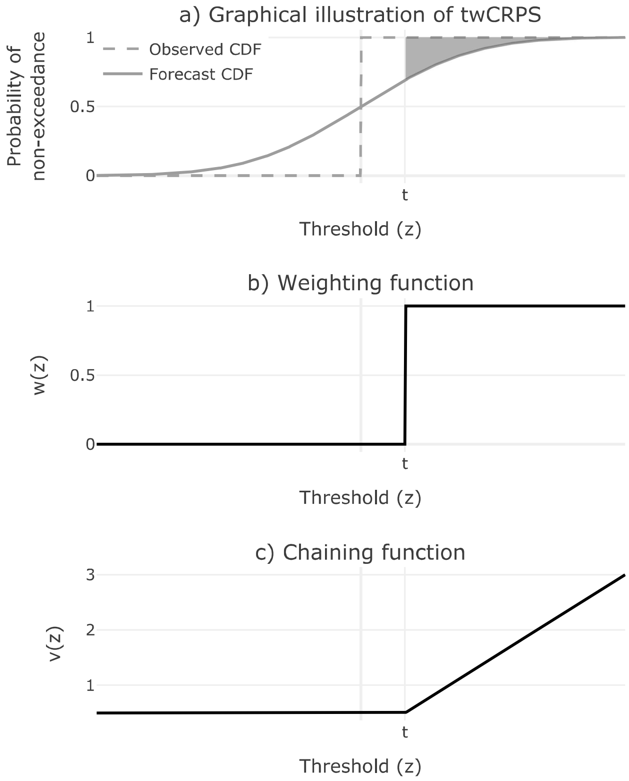

Figure 2: Graphical illustration of the threshold-weighted continuous ranked probability score (twCRPS) with a uniform weight of 1 applied to all thresholds $z > t$ and a weight of 0 applied elsewhere. **(a)** The solid blue curve shows the forecast cumulative distribution function (CDF) and the dashed orange line represents the Heaviside step function of the observation. The twCRPS is the integrated squared difference between the solid blue curve and the dashed orange line, with interval of integration $(x > t)$ indicated by the green shaded region. **(b)** The threshold weight function $w(z) = \mathbb{1}(z > t)$. **(c)** The corresponding chaining function $v(z) = \max(z, t)$.

Recently, Allen et al. (2023) showed that the twCRPS can be adapted for use with ensemble forecasts. The twCRPS for evaluating an ensemble forecast F_{ens} is defined as

$$\text{twCRPS}(F_{ens}, y; v) = \frac{1}{M}\sum_{m=1}^{M} |v(x_m) - v(y)| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{j=1}^{M} |v(x_m) - v(x_j)|, \tag{6}$$

where v is the *chaining function.* The chaining function v is an antiderivative of the threshold weight function $w(z)$, such that

$$v(z) - v(z') = \int_{z'}^{z} w(z)\,\mathrm{d}z. \tag{7}$$

For example, if we wish to assign a threshold weight of 1 to thresholds above a specified threshold t, and a weight of 0 below t, the threshold weight function would be $w(z) = \mathbb{1}(z > t)$, where $\mathbb{1}$ is the indicator function that returns 1 if the condition is true and 0 otherwise. A corresponding chaining function is then $v(z) = \max(z, t)$. Allen et al. (2023); Allen (2024) provide further examples illustrating the relationship between chaining functions and threshold weight functions.

Since we evaluate pseudo-ensembles with varying member sizes, we modify Eq. 6 such that its second term becomes an unbiased estimator, analogous to the "fair" correction applied in Eq. 4 relative to Eq. 3.

$$\text{twCRPS}(F_{ens}, y; v) = \frac{1}{M}\sum_{m=1}^{M} |v(x_m) - v(y)| - \frac{1}{2M(M-1)}\sum_{m=1}^{M}\sum_{j=1}^{M} |v(x_m) - v(x_j)|. \tag{8}$$

3.3 Relating proper scores and spatial methods in practice.

Weather models are used in a wide variety of ways. Here are three examples:

1. Models could be used at each grid point to produce forecasts for a user's location (e.g., via a mobile weather application).

2. A high-resolution ensemble could be used by an operational meteorologist to infer the likelihood of different modes of convection during the afternoon.

3. The dynamic tropopause output of a deterministic model (or single ensemble member) can be used by a decision-support meteorologist to construct a conceptual model of the atmosphere, aiding rapid responses to bespoke queries in an emergency response centre.

These use cases require different evaluation approaches. In the first example, point-to-point evaluation using proper or consistent scoring rules may be appropriate. In the second, point-to-point verification would not capture the spatial realism of convection within the ensemble, necessitating some form of spatial processing. Similarly, in the third case, evaluating the spatial structure of the dynamic tropopause may require a spatial verification method distinct from that used in the second example. These examples illustrate that no single metric can assess all the diverse ways that weather models are used.

It is desirable that forecasts are probabilistic and evaluated using proper scoring rules[1] (Gneiting and Katzfuss, 2014). This is because probabilistic forecasts support optimal decision-making, and proper scores reward honest forecasts while discouraging hedging. Common examples of probabilistic forecasts in meteorology include ensembles, predictive distributions, or binary classifiers (i.e., probability forecasts for binary events). These forecasts are traditionally evaluated point-to-point; however, spatial processing can also be applied to construct probabilistic forecasts. For example, the high-resolution ensemble output

[1] In some cases, only a single-valued forecast is required for the forecast service rather than a full predictive distribution. Consistent scoring functions can then be used to evaluate forecasts expressed through a directive in the form of a statistical functional (Gneiting, 2011); for example, a 90th percentile forecast can be evaluated using a quantile loss.

Table 1: Equivalent neighbourhood sizes for the HRRR and GraphCast-GFS models.

Neighbourhood size (km)	HRRR grid points	GraphCast-GFS grid points
3×3km	1×1	-
21×27km	7×9	1×1
63×81km	21×27	3×3

in the second example could be used to create a probability mass function of various modes of convection, which could then be evaluated using a proper scoring rule.

There have been at least two recent examples of using proper scoring rules to compare single-valued AIWP models against traditional NWP models. First, Brenowitz et al. (2025) created a "lagged" ensemble of single-valued AIWP forecasts that could be evaluated with the CRPS. Secondly, Gneiting et al. (2025) converted single-valued forecasts into a predictive distribution using Isotonic Distributional Regression (IDR) (Henzi et al., 2021) and evaluated the resulting distribution using the CRPS. Both approaches are point-to-point methods that do not account for any spatial information.

Another use case is when meteorologists visually assess a deterministic precipitation forecast around a point to qualitatively estimate the likelihood of various precipitation amounts. A quantitative forecast can be constructed by generating a predictive distribution from a neighbourhood pseudo-ensemble around the observation. Similarly, post-processing techniques often employ neighbourhood approaches on NWP output to generate probabilistic forecasts (Schwartz and Sobash, 2017). The HiRA framework captures this use case by quantifying the benefits of probabilistic forecasts derived from different-resolution models over equivalent spatial areas. Using twCRPS within HiRA allows emphasis on performance across varying decision thresholds.

Pic et al. (2025) sought to formalise several principles in a framework for building interpretable multivariate proper scoring rules that relate to spatial verification methods. In their framework, the construction of neighbourhoods is referred to as a "transformation over patches". In our case, we take a patch, which is the neighbourhood and transform it into a pseudo-ensemble to evaluate with a proper scoring rule.

4 CRPS and twCRPS HiRA results

Our approach is to construct neighbourhood sizes that cover a similar spatial area for both models. Since GraphCast-GFS is on a latitude–longitude projection and its grid cell size varies with latitude, we use a rectangular neighbourhood shape for the HRRR model (with a 3 km grid resolution) that approximates the neighbourhood sizes for GraphCast-GFS over CONUS. This approach differs from most HiRA approaches, which typically use square or circular neighbourhood shapes. The equivalent neighbourhood sizes are shown in Table 1.

Note that, although there is no neighbourhood size for GraphCast-GFS that is equivalent to the 3×3 km point forecast for HRRR, we still evaluate the point-based HRRR to understand the impact of neighbourhood size on the scores because of the double penalty effect. The CRPS of a single ensemble member (i.e., the point-based forecast using a neighbourhood size of 1) is equivalent to the absolute loss. Likewise, threshold-weighted MAE (or threshold-weighted absolute loss) (Taggart, 2021) is equivalent to twCRPS for a single ensemble member.

When aggregating results spatially, we weight the results based on station density by following the approach in Rodwell et al. (2010). This ensures that the verification results are not overly influenced by geographical regions that have a higher concentration of stations compared to others. The station density weights are allowed to vary for each timestep to account for non-continuous reporting from weather stations.

4.1 CRPS results

We first compute mean CRPS results within the HiRA framework to evaluate differences in performance between predicting extreme precipitation and overall forecast performance.

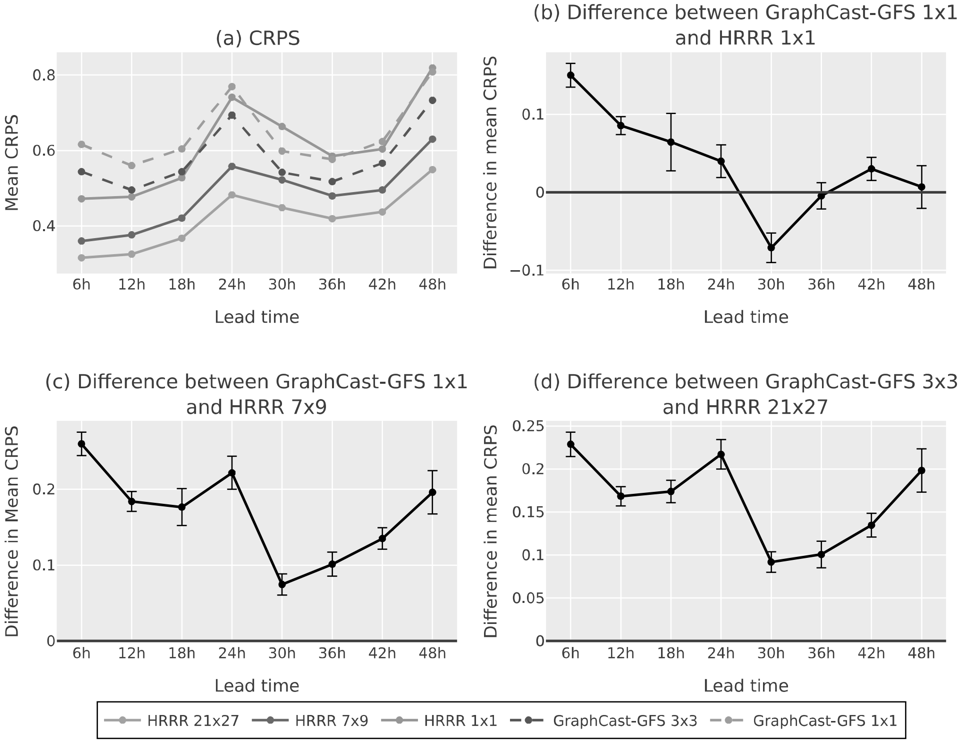

Figure 3: **(a)** Mean CRPS results aggregated across all stations and timesteps. Lower scores are better. **(b)** Difference between GraphCast-GFS 1×1 and HRRR 1×1 with 99% confidence intervals. **(c)** Difference between GraphCast-GFS 1×1 and HRRR 7×9 (21 × 27 km equivalent) with 99% confidence intervals. **(d)** Difference between GraphCast-GFS 3×3 and HRRR 21×27 (63×81 km equivalent)with 99% confidence intervals. In subfigures b-d, positive values indicate that HRRR performed better than GraphCast-GFS for the specified neighbourhoods.

Figure 3a shows the mean CRPS for GraphCast-GFS and HRRR across various neighbourhood sizes. Increasing the neighbourhood size leads to a better score for both models. While one might expect a large improvement in performance when increasing the neighbourhood size in the HRRR as it is a high resolution physical NWP model, an improvement also occurs with GraphCast-GFS despite its smoother appearance. A diurnal trend in model performance can also be seen. When the two models are compared gridpoint-to-gridpoint (Fig. 3b), HRRR outperforms GraphCast-GFS during the first 24 hours, with mixed results at longer lead times. However, when the models are assessed over comparable neighbourhood sizes (Figs. 3c-d), the HRRR demonstrated better performance across all lead times.

4.2 twCRPS results

To assess performance in predicting extremes, we construct a twCRPS weight function that takes the form $w(z) = \mathbb{1}(z > q_\alpha)$, where q_α is the α quantile of the climatological values in the ERA5 dataset for the point at which the ASOS station is located. We set $\alpha = 0.99$ to focus on performance above the climatological 99th percentile. The corresponding chaining function that we use is $v(z) = \max(z, q_\alpha)$. When the mean twCRPS is calculated across time and space, q_α varies by station but is fixed in time. To test the impact of a more extreme threshold, we repeat the computation in Appendix A with $\alpha = 0.999$

to examine performance above the climatological 99.9th percentile.

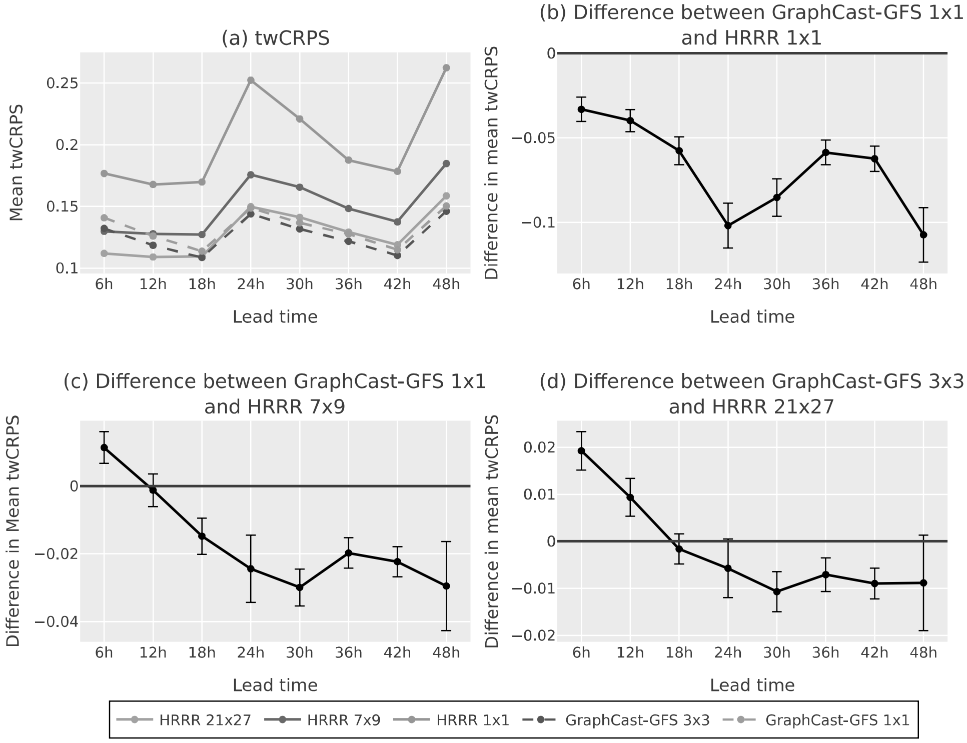

Figure 4: As for Fig. 3 but for the twCRPS with a threshold weight function of $w(z) = \mathbb{1}(z > q_{0.99})$.

In contrast to Fig. 3a, Fig. 4a shows that increasing the neighbourhood size had a larger impact on the performance for the HRRR compared to GraphCast-GFS. When evaluated on a point-to-point basis (Fig. 4b), GraphCast-GFS consistently achieved lower twCRPS scores across all lead times. When evaluated across equivalent neighbohood sizes (Fig. 4c-d), the HRRR performs better at shorter lead times, but not at longer lead times. Potentially, this may be attributed to the HRRR's assimilation of radar data, enhancing its short-term prediction of heavy precipitation. We leave a detailed investigation of this behaviour for future research.

5 Decomposing CRPS across decision thresholds

Since the CRPS is the integral of the Brier score across all thresholds (Eq. 1), we decompose it to visualise the Brier scores across a range of thresholds. This provides us with greater insight within the HiRA framework as to how different models provide value at different decision thresholds. We display results for two lead times in Fig. 5.

Figure 5 shows that the mean CRPS is dominated by non-extreme precipitation amounts (i.e., mean Brier score values from lower thresholds), which is unsurprising given their higher climatological frequency. To account for this, results are split at the 30 mm threshold, and the right-hand panels use a logarithmic vertical axis to better visualise the forecast rankings.

At a 6-hour lead time (top panels), the HRRR model outperforms GraphCast-GFS for lower precipitation thresholds less than 5mm, across all neighbourhood sizes. However, at the 30-hour lead time, this is only true for some neighbourhood sizes. In cases where the mean Brier score curves for HRRR and GraphCast-GFS intersect, the crossing point typically occurs at higher precipitation thresholds for larger HRRR neighbourhood sizes. This may reflect the increased likelihood of double-penalty effects at higher precipitation amounts when using point-to-point verification.

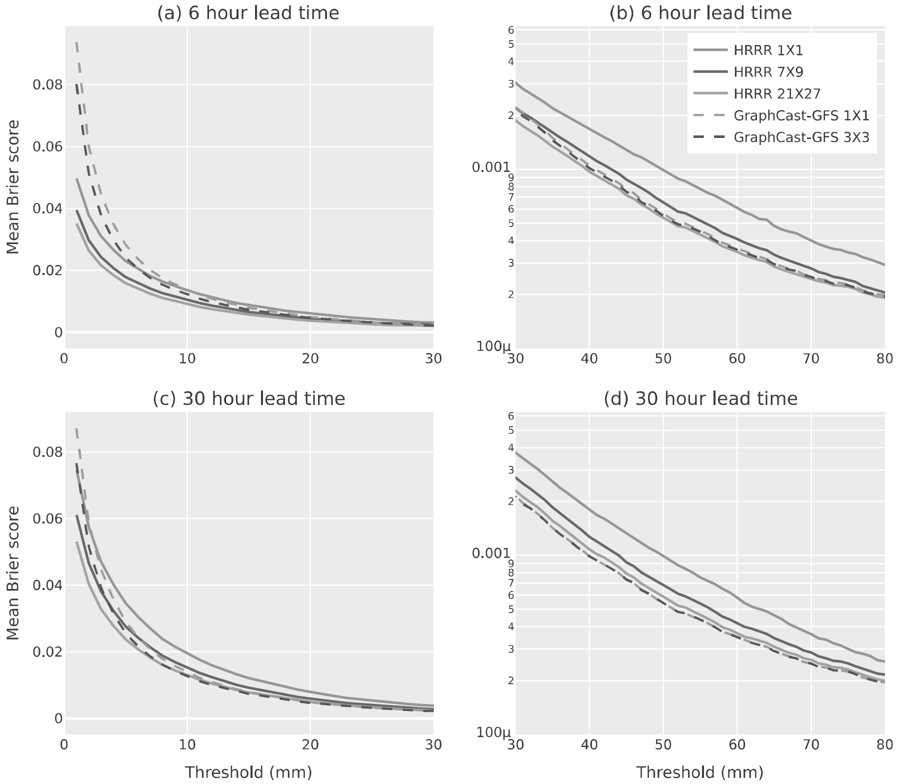

Figure 5: Brier score decomposition of the CRPS within the HiRA framework. Lower scores are better. The left panels **(a)** and **(c)** show the mean Brier score for thresholds below 30 mm, while the right panels **(b)** and **(d)** show the mean Brier score for threshold between 30 and 80 mm with a logarithmic vertical axis. Results are shown for lead time 6 hour forecasts in panels **(a)** and **(b)**, and lead time 30 hour forecasts in **(c)** and **(d)**.

For the upper precipitation thresholds, HRRR 21×27 performs marginally better than GraphCast-GFS 3×3 at the 6-hour lead time, but slightly worse at the 30-hour lead time, which is consistent with the results in Fig. 4.

6 Model climatology

We now assess the agreement between the model climatology and observations to understand any model biases. Quantile-Quantile (Q-Q) plots for the lead time 6-hour and 30-hour point-based forecasts and the observations are shown in Fig. 6. They show that the climatology of HRRR 1×1 closely matches observed climatology of the ASOS observations, while GraphCast-GFS 1×1 is substantially less likely to predict heavier precipitation. This may be partly due to the differences in grid resolution with the HRRR model producing forecasts that match more closely with station-based observations, while GraphCast-GFS produces forecasts that match more closely with a 0.25° grid. This is consistent with the issues with comparing gridded model precipitation forecasts against rain gauge observations highlighted by Tustison et al. (2001).

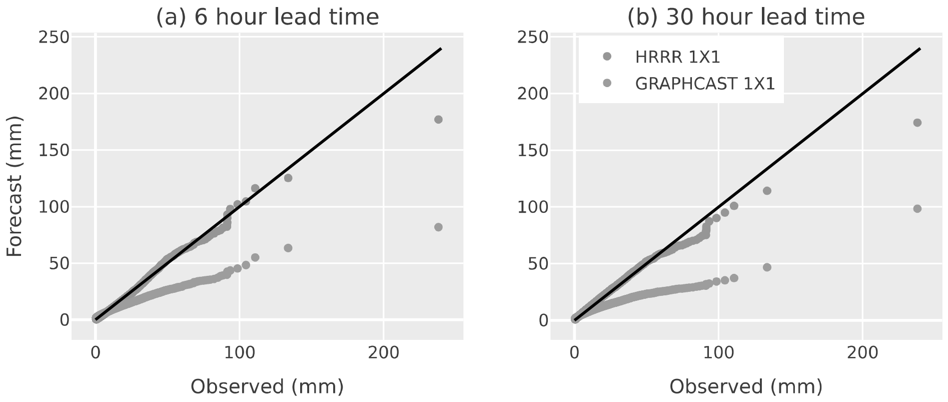

Figure 6: Q-Q plots of observations against forecasts. **(a)** shows results for 6-hour lead time forecasts and **(b)** shows results for 30-hour lead times.

7 Discrimination ability

As demonstrated in the preceding section, the behaviour of GraphCast-GFS may not be consistent with in-situ observations. Nevertheless, it is important to understand the discrimination ability (i.e., potential predictive ability) of the models. This is because meteorologists may be able to learn how the model behaves and use its output accordingly. Alternatively, a model with good discrimination ability could easily be post-processed to correct for any conditional biases. Gneiting et al. (2025) introduced the potential CRPS (PC) to measure and compare the discrimination ability of a deterministic NWP model to that of an AIWP model. This is done by calibrating the model (in sample) using IDR via the EasyUQ method (Walz et al., 2024) to convert the single-valued forecasts to probabilistic forecasts. We adopt a related, but distinct approach. Instead of using IDR on each neighbourhood member, we apply isotonic regression (in sample) to each neighbourhood member. Isotonic regression is a method for fitting a nondecreasing function to a set of forecast-observation pairs (Ayer et al., 1955). We adopt this approach to approximate the behavior of an operational meteorologist who is familiar with typical model biases and applies simple calibrations to the model output. Additionally, it partly circumvents the issue in neighbourhood verification approaches where adjacent grid cells in the neighbourhood are not representative of the station site (e.g., in areas of varying topography). This approach, however, is unlikely to be as effective as IDR in producing a well calibrated predictive distribution if neighbourhood members were used as covariates.

We then take a CORP-like (Consistent, Optimally binned, Reproducible, and Pool-Adjacent-Violators (PAV) algorithm-based) decomposition approach (Dimitriadis et al., 2021; Arnold et al., 2024) to the twCPRS score. The CORP decomposition approach produces three compoents: discrimination, miscalibration, and uncertainty. Here, we only focus on the discrimination (DSC) component which we define as

$$\mathrm{DSC} = \overline{\mathrm{twCRPS}}_C - \overline{\mathrm{twCRPS}}_R, \tag{9}$$

where $\overline{\mathrm{twCRPS}}_C$ is the mean twCRPS score of a sample climatological forecast at each station and $\overline{\mathrm{twCRPS}}_R$ is the mean twCRPS score of the (re)calibrated forecasts. Positive DSC values indicate that there is discrimination ability and larger values indicate more discrimination ability. We display the DSC results in Fig. 7.

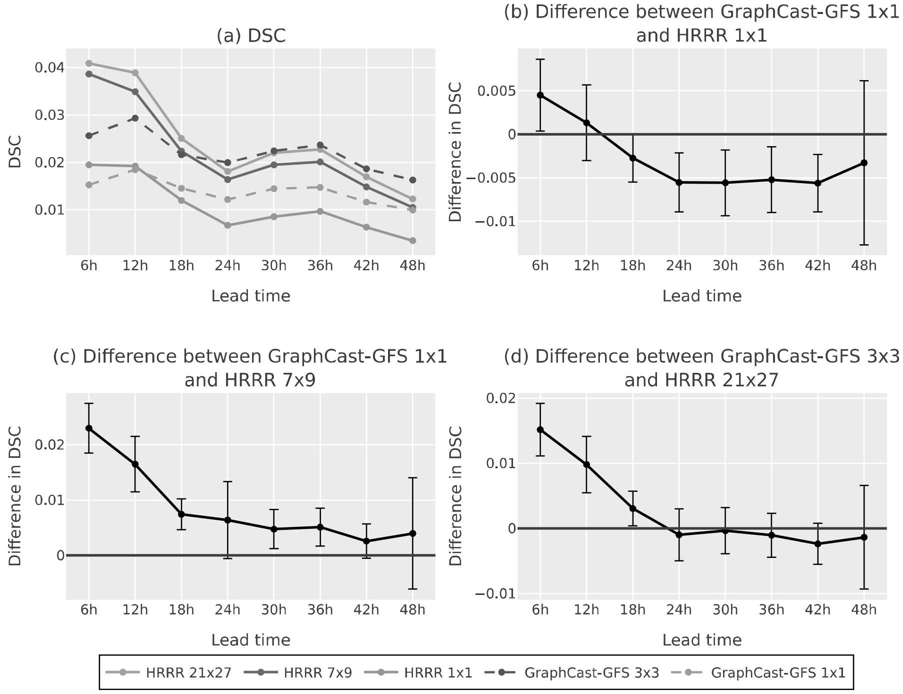

Figure 7: **(a)** DSC results. Higher values indicate more discrimination ability for predicting extremes (all thresholds above the climatological 99th percentile). **(b)** Difference between GraphCast-GFS 1×1 and HRRR 1×1 with 99% confidence intervals. **(c)** Difference between GraphCast-GFS 1×1 and HRRR 7×9 with 99% confidence intervals. **(d)** Difference between GraphCast-GFS 3×3 and HRRR 21×27 with 99% confidence intervals. In subfigures b-d, positive values indicate that HRRR had more discrimination ability than GraphCast-GFS for the specified neighbourhoods.

The DSC results, which measures discrimination ability, show that while increasing neighbourhood size leads to better discrimination ability, the HRRR and GraphCast-GFS are sometimes ranked differently compared to Fig. 4 measures overall predictive performance rather than discrimination ability.

For all equivalent neighbourhood sizes, the HRRR model has superior discrimination ability at a lead time of 6 hours, but this difference decreases with longer lead times and GraphCast-GFS has more discrimination ability for lead times 24 hours and beyond for the GraphCast-GFS 1×1 vs HRRR 1×1 comparison and the GraphCast-GFS 3×3 vs HRRR 21×27 comparison. Differences in discrimination ability were not significant for the latter comparison.

8 Software and future collaboration on verification tools between DTC and the Bureau of Meteorology

As part of this visit, Python tools for calculating the twCRPS for ensembles were made open source as part of the *scores* Python package (Leeuwenburg et al., 2024) since twCRPS is not available in METplus (Brown et al., 2021). Three functions were added to the *scores* package:

- `scores.probability.tail_tw_crps_for_ensemble` for emphasising performance for an upper or lower tail (e.g., for extremes)
- `scores.probability.interval_tw_crps_for_ensemble` for emphasising performance across an interval of decision thresholds
- `scores.probability.tw_crps_for_ensemble` for emphasising performance based on custom weights derived from a provided chaining function.

This implementation of twCRPS allows the calculation of N-dimensional data (e.g., gridded or site-based) with xarray. It is designed to scale with Dask if needed and has 100% unit test coverage.

A tutorial was developed to demonstrate how twCRPS can be used in a point-to-point (non-HiRA) context `https://scores.readthedocs.io/en/stable/tutorials/Threshold_Weighted_CRPS_for_Ensembles.html`. The tutorial includes examples that compare the ECMWF IFS ensemble (physical NWP model) against the Neural GCM (hybrid model) ensemble. Additionally, the fair Brier score was added to the *scores* package as `scores.probability.brier_score_for_ensemble` which was used to generate the results in Fig. 5.

During the visit, we discussed the possibility of integrating *scores* into METplus. Currently, MET does most of its core verification calculations in C++ which makes it difficult for the community to contribute new methods to MET. In contrast, *scores* is written 100% in Python. By allowing MET to call *scores* for verification calculations, there would be several advantages:

1. It would provide MET with access to a wide range of verification methods that are already implemented in *scores* that are not currently available in MET (e.g., threshold weighted scores, block-bootstrapping).
2. It would allow the community to contribute new verification methods for MET without needing to write C++ code. It would open up a clear pathway for Bureau of Meteorology researchers to contribute new methods that can be used operationally in METplus.
3. It would open up the possibility of using Dask to scale calculations across multiple cores or nodes if required.
4. Future spatial verification methods may use Python-based image processing libraries. This would provide a pathway for MET to use these methods.

A GitHub issue has been opened to discuss this possibility `https://github.com/dtcenter/MET/issues/2978`. This would provide a clear pathway for integrating the work from this visit into METplus.

9 Conclusions and future research

This report demonstrated an approach to evaluating how an AIWP model compares to a high-resolution physical NWP model in predicting climatologically extreme 6-hourly precipitation. It combines two existing techniques: the HiRA framework and twCRPS. Model performance was assessed within a framework representative of potential operational use by meteorologists or simple post-processing systems. As with all NWP verification methods, this approach does not definitively determine which model is superior across all use cases.

When models were compared using equivalent neighbourhood sizes, HRRR consistently outperformed GraphCast-GFS across all lead times as measured by CRPS. However, when focusing on predictive performance of extreme precipitation, HRRR only outperformed GraphCast-GFS at short lead times.

Decomposing CRPS by decision threshold provided further insight into which thresholds each model handled more effectively. Where Brier score curves for HRRR and GraphCast-GFS intersected, this generally occurred at higher precipitation thresholds for larger HRRR neighbourhood sizes.

The approach was extended to measure discrimination ability of the models to predict extreme precipitation. This was important to measure since GraphCast-GFS forecasts of heavy precipitation showed an under-forecast bias when evaluated against rain gauge data. Increasing the neighbourhood size led to superior discrimination ability. When GraphCast-GFS 3×3 vs HRRR 21×27 were compared, the high resolution model had superior discrimination ability at short lead times, but not beyond 24 hours.

This verification approach has several strengths:

1. It is a spatial method that can address the double penalty effect within a specified distance.

2. It supports intercomparison of models with different resolutions without requiring re-gridding.

3. It can be used to evaluate models against in-situ observations.

4. It can be threshold-weighted to focus on extremes or other important decision thresholds.

5. It uses proper scoring rules within a user-focused framework, avoiding selection bias associated with conditioning on extreme observed or forecast events.

6. It supports aggregation across domains by accounting for climatological differences via threshold weighting if required.

As noted by Pic et al. (2025), there is an opportunity to bridge the gap between the spatial verification methods and proper scoring rules communities. This work offers one such integration, and raises several opportunities for future research:

- The approach could be extended to evaluate multivariate forecasts by using threshold weighted multivariate scores, such as the threshold-weighted variogram score (Allen et al., 2023) which is a proper scoring rule. This approach could be used to evaluate compound events as well as different variables simultaneously (e.g., surface wind speed, temeprature and relative humidity for fire weather).

- Comparing this approach to other spatial verification methods.

- Applying threshold-weighted scoring rules within HiRA to ensemble forecasts, as in Mittermaier and Csima (2017).

- Extending the approach to also account for timing errors.

- Applying neighbourhood twCRPS to forecasts evaluated with gridded rainfall observations, which is similar to Stein and Stoop (2022) who used a neighbourhood CRPS with gridded data.

- Comparing a potential CRPS measure (Gneiting et al., 2025) that applies IDR using neighbourhood members as covariates. This may be useful when models are spatially sharper.

Finally, while most prior comparisons between AIWP and traditional NWP models have relied on point-to-point verification, this study shows that model rankings can differ when assessed from a spatial perspective, consistent with findings from Radford et al. (2025a). If AIWP models are to be operationalised, adopting spatial verification methods aligned with their practical use will be essential for accurately understanding their performance.

Acknowledgements

Thank you to DTC for hosting me as a visiting scientist. Thank you to Tracy Hertneky, Molly Smith, Michael Ek, John Halley Gotway, Kathryn Neman, Jennifer Bolton, Barabara Brown, Michelle Harrold, George McGabe, Minna Win-Gildenmeister, Louisa Nance, Jeffrey Hamilton, and Weiwei Li for supporting my visit and ensuring that I had a wonderful trip to Boulder.

A Appendix A: twCRPS results for the 0.999 climatological threshold

Results for the twCRPS using a 0.999 climatological threshold are shown in Fig. 8. The only instance that the HRRR was significantly better than GraphCast-GFS was the 6-hour lead time when GraphCast-GFS 3×3 and HRRR 21×27 were compared. Results were also calculated using a fixed 50mm threshold at all locations and were similar (not shown).

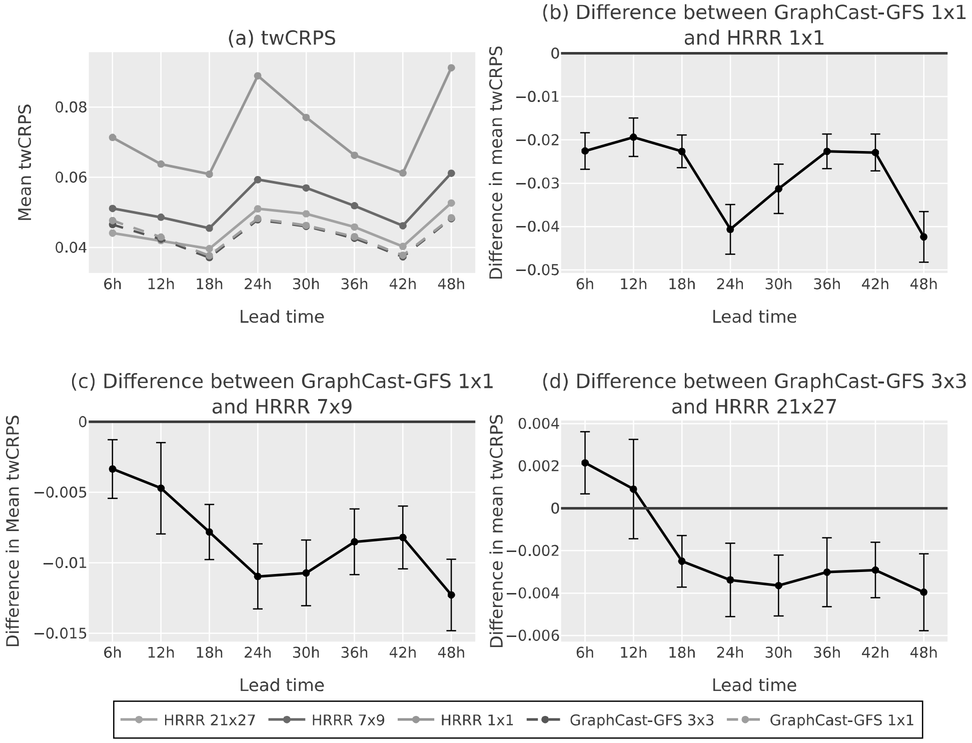

Figure 8: As for Fig. 3 but for the twCRPS with a threshold weight function of $w(z) = \mathbb{1}(z > q_{0.999})$.

References

Abdi, D., Jankov, I., Madden, P., Vargas, V., Smith, T. A., Frolov, S., Flora, M., and Potvin, C.: HRRRCast: a data-driven emulator for regional weather forecasting at convection allowing scales, https://arxiv.org/abs/2507.05658, 2025.

Adamov, S., Oskarsson, J., Denby, L., Landelius, T., Hintz, K., Christiansen, S., Schicker, I., Osuna, C., Lindsten, F., Fuhrer, O., and Schemm, S.: Building Machine Learning Limited Area Models: Kilometer-Scale Weather Forecasting in Realistic Settings, https://arxiv.org/abs/2504.09340, 2025.

Allen, S.: Weighted scoringRules: Emphasizing Particular Outcomes When Evaluating Probabilistic Forecasts, Journal of Statistical Software, 110, https://doi.org/10.18637/jss.v110.i08, 2024.

Allen, S., Ginsbourger, D., and Ziegel, J.: Evaluating Forecasts for High-Impact Events Using Transformed Kernel Scores, SIAM/ASA Journal on Uncertainty Quantification, 11, 906–940, https://doi.org/10.1137/22m1532184, 2023.

Arnold, S., Walz, E.-M., Ziegel, J., and Gneiting, T.: Decompositions of the mean continuous ranked probability score, Electronic Journal of Statistics, 18, https://doi.org/10.1214/24-ejs2316, 2024.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E.: An Empirical Distribution Function for Sampling with Incomplete Information, The Annals of Mathematical Statistics, 26, 641–647, https://doi.org/10.1214/aoms/1177728423, 1955.

Baringhaus, L. and Franz, C.: On a new multivariate two-sample test, Journal of Multivariate Analysis, 88, 190–206, https://doi.org/10.1016/s0047-259x(03)00079-4, 2004.

Ben Bouallègue, Z., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T. K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., and Pappenberger, F.: The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context, Bulletin of the American Meteorological Society, 105, E864–E883, https://doi.org/10.1175/bams-d-23-0162.1, 2024.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 619, 533–538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Brenowitz, N. D., Cohen, Y., Pathak, J., Mahesh, A., Bonev, B., Kurth, T., Durran, D. R., Harrington, P., and Pritchard, M. S.: A Practical Probabilistic Benchmark for AI Weather Models, Geophysical Research Letters, 52, https://doi.org/10.1029/2024gl113656, 2025.

Brier, G. W.: VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY, Monthly Weather Review, 78, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2, 1950.

Brown, B., Jensen, T., Gotway, J. H., Bullock, R., Gilleland, E., Fowler, T., Newman, K., Adriaansen, D., Blank, L., Burek, T., Harrold, M., Hertneky, T., Kalb, C., Kucera, P., Nance, L., Opatz, J., Vigh, J., and Wolff, J.: The Model Evaluation Tools (MET): More than a Decade of Community-Supported Forecast Verification, Bulletin of the American Meteorological Society, 102, E782–E807, https://doi.org/10.1175/bams-d-19-0093.1, 2021.

Charlton-Perez, A. J., Dacre, H. F., Driscoll, S., Gray, S. L., Harvey, B., Harvey, N. J., Hunt, K. M. R., Lee, R. W., Swaminathan, R., Vandaele, R., and Volonté, A.: Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán, npj Climate and Atmospheric Science, 7, https://doi.org/10.1038/s41612-024-00638-w, 2024.

Crocker, R., Maksymczuk, J., Mittermaier, M., Tonani, M., and Pequignet, C.: An approach to the verification of high-resolution ocean models using spatial methods, Ocean Science, 16, 831–845, https://doi.org/10.5194/os-16-831-2020, 2020.

Dimitriadis, T., Gneiting, T., and Jordan, A. I.: Stable reliability diagrams for probabilistic classifiers, Proceedings of the National Academy of Sciences, 118, https://doi.org/10.1073/pnas.2016191118, 2021.

Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M. P., Ebert, E. E., Brown, B. G., and Wilson, L. J.: The Setup of the MesoVICT Project, Bulletin of the American Meteorological Society, 99, 1887–1906, https://doi.org/10.1175/bams-d-17-0164.1, 2018.

Dowell, D. C., Alexander, C. R., James, E. P., Weygandt, S. S., Benjamin, S. G., Manikin, G. S., Blake, B. T., Brown, J. M., Olson, J. B., Hu, M., Smirnova, T. G., Ladwig, T., Kenyon, J. S., Ahmadov, R., Turner, D. D., Duda, J. D., and Alcott, T. I.: The High-Resolution Rapid Refresh (HRRR): An Hourly Updating Convection-Allowing Forecast Model. Part I: Motivation and System Description, Weather and Forecasting, 37, 1371–1395, https://doi.org/10.1175/waf-d-21-0151.1, 2022.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorological Applications, 15, 51–64, https://doi.org/10.1002/met.25, 2008.

Ebert, E. E.: Neighborhood Verification: A Strategy for Rewarding Close Forecasts, Weather and Forecasting, 24, 1498–1510, https://doi.org/10.1175/2009waf2222251.1, 2009.

Epstein, E. S.: A Scoring System for Probability Forecasts of Ranked Categories, Journal of Applied Meteorology, 8, 985–987, https://doi.org/10.1175/1520-0450(1969)008<0985:assfpf>2.0.co;2, 1969.

Ferro, C. A. T.: Fair scores for ensemble forecasts: Fair Scores for Ensemble Forecasts, Quarterly Journal of the Royal Meteorological Society, 140, 1917–1923, https://doi.org/10.1002/qj.2270, 2013.

Fovell, R. G. and Gallagher, A.: An Evaluation of Surface Wind and Gust Forecasts from the High-Resolution Rapid Refresh Model, Weather and Forecasting, 37, 1049–1068, https://doi.org/10.1175/waf-d-21-0176.1, 2022.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of Spatial Forecast Verification Methods, Weather and Forecasting, 24, 1416–1430, https://doi.org/10.1175/2009waf2222269.1, 2009.

Gneiting, T.: Making and Evaluating Point Forecasts, Journal of the American Statistical Association, 106, 746–762, https://doi.org/10.1198/jasa.2011.r10138, 2011.

Gneiting, T. and Katzfuss, M.: Probabilistic Forecasting, Annual Review of Statistics and Its Application, 1, 125–151, https://doi.org/https://doi.org/10.1146/annurev-statistics-062713-085831, 2014.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association, 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Gneiting, T. and Ranjan, R.: Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules, Journal of Business & Economic Statistics, 29, 411–422, https://doi.org/10.1198/jbes.2010.08110, 2011.

Gneiting, T., Biegert, T., Kraus, K., Walz, E.-M., Jordan, A. I., and Lerch, S.: Probabilistic measures afford fair comparisons of AIWP and NWP model output, `https://arxiv.org/abs/2506.03744`, 2025.

Henzi, A., Ziegel, J. F., and Gneiting, T.: Isotonic Distributional Regression, Journal of the Royal Statistical Society Series B: Statistical Methodology, 83, 963–993, https://doi.org/10.1111/rssb.12450, 2021.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly journal of the royal meteorological society, 146, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, 2020.

Ikeda, K., Steiner, M., Pinto, J., and Alexander, C.: Evaluation of Cold-Season Precipitation Forecasts Generated by the Hourly Updating High-Resolution Rapid Refresh Model, Weather and Forecasting, 28, 921–939, https://doi.org/10.1175/waf-d-12-00085.1, 2013.

Jin, W., Weyn, J., Zhao, P., Xiang, S., Bian, J., Fang, Z., Dong, H., Sun, H., Thambiratnam, K., and Zhang, Q.: WeatherReal: A Benchmark Based on In-Situ Observations for Evaluating Weather Models, `https://arxiv.org/abs/2409.09371`, 2024.

Keisler, R.: Forecasting Global Weather with Graph Neural Networks, `https://arxiv.org/abs/2202.07575`, 2022.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C. A., Lessig, C., Maier-Gerber, M., Magnusson, L., Bouallègue, Z. B., Nemesio, A. P., Dueben, P. D., Brown, A., Pappenberger, F., and Rabier, F.: AIFS – ECMWF's data-driven forecasting system, `https://arxiv.org/abs/2406.01465`, 2024a.

Lang, S., Alexe, M., Clare, M. C. A., Roberts, C., Adewoyin, R., Bouallègue, Z. B., Chantry, M., Dramsch, J., Dueben, P. D., Hahner, S., Maciel, P., Prieto-Nemesio, A., O'Brien, C., Pinault, F., Polster, J., Raoult, B., Tietsche, S., and Leutbecher, M.: AIFS-CRPS: Ensemble forecasting using a model trained with a loss function based on the Continuous Ranked Probability Score, `https://arxiv.org/abs/2412.15832`, 2024b.

Lavers, D. A., Simmons, A., Vamborg, F., and Rodwell, M. J.: An evaluation of ERA5 precipitation for climate monitoring, Quarterly Journal of the Royal Meteorological Society, 148, 3152–3165, https://doi.org/10.1002/qj.4351, 2022.

Leeuwenburg, T., Loveday, N., Ebert, E. E., Cook, H., Khanarmuei, M., Taggart, R. J., Ramanathan, N., Carroll, M., Chong, S., Griffiths, A., and Sharples, J.: scores: A Python package for verifying and evaluating models and predictions with xarray, Journal of Open Source Software, 9, 6889, https://doi.org/10.21105/joss.06889, 2024.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T.: Forecaster's Dilemma: Extreme Events and Forecast Evaluation, Statistical Science, 32, https://doi.org/10.1214/16-sts588, 2017.

Loveday, N., Griffiths, D., Leeuwenburg, T., Taggart, R. J., Pagano, T. C., Cheng, G., Plastow, K., Ebert, E. E., Templeton, C., Carroll, M., Khanarmuei, M., and Nagpal, I.: The Jive Verification System and Its Transformative Impact on Weather Forecasting Operations, Bulletin of the American Meteorological Society, 105, E2047–E2063, https://doi.org/10.1175/bams-d-23-0267.1, 2024.

Mass, C. F., Ovens, D., Westrick, K., and Colle, B. A.: Does Increasing Horizontal Resolution Produce More Skillful Forecasts?, Bulletin of the American Meteorological Society, 83, 407–430, https://doi.org/10.1175/1520-0477(2002)083<0407:dihrpm>2.3.co;2, 2002.

Matheson, J. E. and Winkler, R. L.: Scoring Rules for Continuous Probability Distributions, Management Science, 22, 1087–1096, https://doi.org/10.1287/mnsc.22.10.1087, 1976.

Mittermaier, M. P.: A Strategy for Verifying Near-Convection-Resolving Model Forecasts at Observing Sites, Weather and Forecasting, 29, 185–204, https://doi.org/10.1175/waf-d-12-00075.1, 2014.

Mittermaier, M. P. and Csima, G.: Ensemble versus Deterministic Performance at the Kilometer Scale, Weather and Forecasting, 32, 1697–1709, https://doi.org/10.1175/waf-d-16-0164.1, 2017.

Morisseau, H., Zhu, H., Hudson, D., and de Burgh-Day, C.: Object-oriented verification of TC-Jasper rainfall forecasts: Machine learning, http://www.bom.gov.au/research/publications/researchreports/BRR-106.pdf, 2025.

Nipen, T. N., Haugen, H. H., Ingstad, M. S., Nordhagen, E. M., Salihi, A. F. S., Tedesco, P., Seierstad, I. A., Kristiansen, J., Lang, S., Alexe, M., Dramsch, J., Raoult, B., Mertes, G., and Chantry, M.: Regional data-driven weather modeling with a global stretched-grid, https://arxiv.org/abs/2409.02891, 2024.

NWS: Automated Surface Observing System (ASOS) user's guide, Tech. rep., NOAA, 1998.

Olivetti, L. and Messori, G.: Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather, and GraphCast, Geoscientific Model Development, 17, 7915–7962, https://doi.org/10.5194/gmd-17-7915-2024, 2024.

Pagano, T. C., Casati, B., Landman, S., Loveday, N., Taggart, R., Ebert, E. E., Khanarmuei, M., Jensen, T. L., Mittermaier, M., Roberts, H., Willington, S., Roberts, N., Sowko, M., Strassberg, G., Kluepfel, C., Bullock, T. A., Turner, D. D., Pappenberger, F., Osborne, N., and Noble, C.: Challenges of Operational Weather Forecast Verification and Evaluation, Bulletin of the American Meteorological Society, 105, E789–E802, https://doi.org/10.1175/bams-d-22-0257.1, 2024.

Pic, R., Dombry, C., Naveau, P., and Taillardat, M.: Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation, Advances in Statistical Climatology, Meteorology and Oceanography, 11, 23–58, https://doi.org/10.5194/ascmo-11-23-2025, 2025.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., and Willson, M.: Probabilistic weather forecasting with machine learning, Nature, 637, 84–90, https://doi.org/10.1038/s41586-024-08252-9, 2024.

Radford, J. T., Ebert-Uphoff, I., and Stewart, J. Q.: A Comparison of AI Weather Prediction and Numerical Weather Prediction Models for 1–7-Day Precipitation Forecasts, Weather and Forecasting, https://doi.org/10.1175/waf-d-24-0081.1, 2025a.

Radford, J. T., Ebert-Uphoff, I., Stewart, J. Q., Musgrave, K. D., DeMaria, R., Tourville, N., and Hilburn, K.: Accelerating Community-Wide Evaluation of AI Models for Global Weather Prediction by Facilitating Access to Model Output, Bulletin of the American Meteorological Society, 106, E68–E76, https://doi.org/10.1175/bams-d-24-0057.1, 2025b.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallegue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F.: WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models, Journal of Advances in Modeling Earth Systems, 16, https://doi.org/10.1029/2023ms004019, 2024.

Rodwell, M. J., Richardson, D. S., Hewson, T. D., and Haiden, T.: A new equitable score suitable for verifying precipitation in numerical weather prediction, Quarterly Journal of the Royal Meteorological Society, 136, 1344–1363, https://doi.org/10.1002/qj.656, 2010.

Schwartz, C. S. and Sobash, R. A.: Generating Probabilistic Forecasts from Convection-Allowing Ensembles Using Neighborhood Approaches: A Review and Recommendations, Monthly Weather Review, 145, 3397–3418, https://doi.org/10.1175/mwr-d-16-0400.1, 2017.

Stein, J. and Stoop, F.: Neighborhood-Based Ensemble Evaluation Using the CRPS, Monthly Weather Review, 150, 1901–1914, https://doi.org/10.1175/mwr-d-21-0224.1, 2022.

Subich, C., Husain, S. Z., Separovic, L., and Yang, J.: Fixing the Double Penalty in Data-Driven Weather Forecasting Through a Modified Spherical Harmonic Loss Function, https://arxiv.org/abs/2501.19374, 2025.

Székely, G. J. and Rizzo, M. L.: A new test for multivariate normality, Journal of Multivariate Analysis, 93, 58–80, https://doi.org/10.1016/j.jmva.2003.12.002, 2005.

Taggart, R.: Evaluation of point forecasts for extreme events using consistent scoring functions, Quarterly Journal of the Royal Meteorological Society, 148, 306–320, https://doi.org/10.1002/qj.4206, 2021.

Tustison, B., Harris, D., and Foufoula-Georgiou, E.: Scale issues in verification of precipitation forecasts, Journal of Geophysical Research: Atmospheres, 106, 11 775–11 784, https://doi.org/10.1029/2001jd900066, 2001.

Walz, E.-M., Henzi, A., Ziegel, J., and Gneiting, T.: Easy Uncertainty Quantification (EasyUQ): Generating Predictive Distributions from Single-Valued Model Output, SIAM Review, 66, 91–122, https://doi.org/10.1137/22m1541915, 2024.

Wessel, J. B., Ferro, C. A. T., Evans, G. R., and Kwasniok, F.: Improving probabilistic forecasts of extreme wind speeds by training statistical post-processing models with weighted scoring rules, Monthly Weather Review, https://doi.org/10.1175/mwr-d-24-0151.1, 2025.

Winkler, R. L. and Murphy, A. H.: “Good” Probability Assessors, Journal of Applied Meteorology, 7, 751–758, https://doi.org/10.1175/1520-0450(1968)007<0751:pa>2.0.co;2, 1968.

WMO: Guide to Hydrological Practices, Tech. Rep. 168, World Meteorological Organization, 1994.

Zhang, Z., Fischer, E., Zscheischler, J., and Engelke, S.: Numerical models outperform AI weather forecasts of record-breaking extremes, https://arxiv.org/abs/2508.15724, 2025.