Australian Government
Bureau of Meteorology

Machine-learned forecasting and post-processing: probabilistic calibration of AIFS forecasts with RainForests

Felix Esperson, Belinda Trotta

February 2026

Bureau Research Report - 122

Machine-learned forecasting and post-processing: probabilistic calibration of AIFS forecasts with RainForests

Felix Esperson, Belinda Trotta

Bureau of Meteorology

Bureau Research Report No. 122

February 2026

National Library of Australia Cataloguing-in-Publication entry

Authors: Felix Esperson, Belinda Trotta

Title : Machine-learned forecasting and post-processing: probabilistic calibration of AIFS forecasts with RainForests

ISBN : 978-1-923469-14-3

ISSN : 2206-3366

Series: Bureau Research Report - BRR 122

Enquiries should be addressed to:

Belinda Trotta:

Bureau of Meteorology
GPO Box 1289, Melbourne
Victoria 3001, Australia

belinda.trotta@bom.gov.au

Copyright and Disclaimer

Contents

List of Figures

List of Tables

Executive summary

Rainfall forecasts typically undergo post-processing in order to improve their accuracy and usability relative to direct model output. Promising machine-learning (ML) based methods for both rainfall forecasting and post-processing have been developed in recent years. However, there has been little exploration into the interaction between ML forecasts and ML post-processing methods. In this study, we evaluate the application of the RainForests post-processing system with the Artificial Intelligence Forecasting System (AIFS) developed by the European Centre for Medium-Range Weather Forecasts (ECMWF). RainForests has been shown to improve traditional physics-based numerical weather forecasts. Using ECMWF's physics-based High-Resolution model (HRES) as a benchmark, we found that post-processed AIFS consistently outperforms post-processed HRES across key metrics including CRPS, RMSE and Brier scores. Qualitative analysis of extreme rainfall forecasts shows that while calibration suppresses peak rainfall in both systems, AIFS is less strongly dampened than HRES and better preserves the spatial structure and intensity of high-impact events. These results indicate that RainForests post-processing is effective when applied to AIFS forecasts and support the inclusion of ML-based forecasting in the Bureau of Meteorology's operational system.

1 Introduction

Recent developments in machine learning (ML) have enabled new approaches to both weather forecasting and forecast post-processing (de Burgh-Day and Leeuwenburg, 2023; Trotta et al., 2024b). ML forecasting models are trained on large historical datasets and are optimised with respect to a loss function; for deterministic forecasts this is often a form of mean squared error (MSE) (Van Poecke et al., 2025). As noted by Moldovan et al. (2025), ML forecasts trained to minimise MSE tend to be more spatially smooth than those produced by traditional physics-based numerical weather prediction (NWP) models, due to the "double-penalty" effect and the fact that MSE especially penalises large errors (Moldovan et al., 2025). While these smoother forecasts are generally less realistic looking, they are often less error-prone at longer lead-times.

In operational settings, forecast post-processing is typically applied to improve forecast quality and refine the model output for its intended use. For example, post-processing models trained on site observations can be used to downscale NWP forecasts from the grid to the point scale. Post-processing can also incorporate additional information, such as forecasts of related variables, to improve the accuracy of the output. Vannitsem et al., 2021 gives an overview of the reasons post-processing is necessary, and the kinds of techniques commonly used. The Bureau of Meteorology (herein referred to as the Bureau) currently post-processes its operational rainfall forecast using an ML-based system called RainForests (Trotta et al., 2024a). RainForests has been shown to substantially improve forecast quality for physics-based NWP models (Trotta et al., 2024b). Until now, RainForests has only been tested with NWP forecasts so it is not known whether similar post-processing improvements could be found for ML-based models.

The European Centre for Medium-Range Weather Forecasts (ECMWF) has recently developed the ML-based Artificial Intelligence Forecasting System (AIFS). We consider the deterministic version of AIFS, which is trained to minimise area-weighted MSE (Lang et al., 2024). In the paper just cited, Lang et. al. benchmark AIFS against ECMWF's physics-based High-Resolution model (HRES), and show that many AIFS forecast variables exhibit comparable skill to HRES and, at longer lead-times, often outperform it. HRES currently contributes to the Bureau's operational precipitation forecasting system, alongside other deterministic and ensemble models (Trotta et al., 2024a). As a deterministic global system, HRES can be post-processed in a manner closely analogous to AIFS, making it a natural and operationally relevant benchmark.

At present, there has been limited investigation into the combined use of ML-based forecasting systems with ML-based post-processing (Trotta et al., 2025). RainForests is already operational at the Bureau, while AIFS displays strong performance as a global forecasting system. It is therefore important to evaluate whether RainForests can effectively post-

process AIFS, enabling its consideration in the operational forecast blend. Recent work by Trotta et al. (2025) demonstrates that statistical post-processing can improve surface-based AIFS variables, indicating possible benefits for rainfall as well.

Traditional physics-based models produce deterministic forecasts by simulating a physically plausible evolution of the atmosphere from the best available estimate of its current state. To obtain a probabilistic forecast, post-processing methods are then applied to derive statistically well-calibrated probability distributions from this deterministic output. In contrast to physics-based models like HRES, AIFS is directly optimised for quantitative accuracy. RainForests is likewise trained for quantitative accuracy, but for probabilistic rather than deterministic outputs. This raises the question, which we address in this study, of whether using RainForests to post-process AIFS will deliver a similar level of improvement as it does for HRES.

2 Methods

Due to time and technical constraints, the Bureau's full operational forecasting system cannot be evaluated directly in this study; therefore, HRES is used as a simpler proxy. Because RainForests has already been extensively validated when applied to physics-based models such as HRES, comparable or superior performance from AIFS when combined with RainForests would provide strong evidence supporting the integration of AIFS into the Bureau's operational rainfall forecast.

We begin by pre-processing AIFS and HRES forecasts into a consistent format suitable for training ML models. RainForests calibration models are then trained using a subset of the available data. These models are subsequently applied to post-process raw AIFS and HRES forecasts using the calibration methods from the original RainForests study (Trotta et al., 2024b). We also investigate altering the original RainForests training methodology by training models for combined lead-times instead of training a unique model for each lead-time. Additionally, we make use of the Optuna package (Akiba et al., 2019) in an attempt to optimise model training parameters.

2.1 Data

The data used in this study consists of historical forecasts from two operational ECMWF models, AIFS and HRES, as well as rainfall data from the Bureau. The overlapping period common to all datasets includes forecast base times between 1 March 2024 to 31 July 2025. An overview of the data pre-processing workflow is provided in Figure 1. A repository for code used for this study can be found at https://github.com/forecast-improvement/aifs-rainforests-paper-code.

RainForests post-processing was implemented using the IMPROVER framework, which provides a consistent data model and reproducible implementation of the RainForests methodology (Roberts et al., 2023; Trotta et al., 2024b). Using IMPROVER ensured that both AIFS and HRES forecasts were post-processed in a manner that simulates the Bureau's operational forecast.

The set of input NWP variables used for RainForests differed slightly between the two forecast models, reflecting differences in model output availability. Both models use forecasts of total and convective precipitation and windspeed. HRES also includes an additional variable for Convective Available Potential Energy (CAPE), which is not available in AIFS. In addition to these NWP forecast variables, both models include, as a feature, modelled clear-sky solar radiation, which serves as a proxy for both time of year and latitude. Further details about the model features can be found in Section 2.3.1.

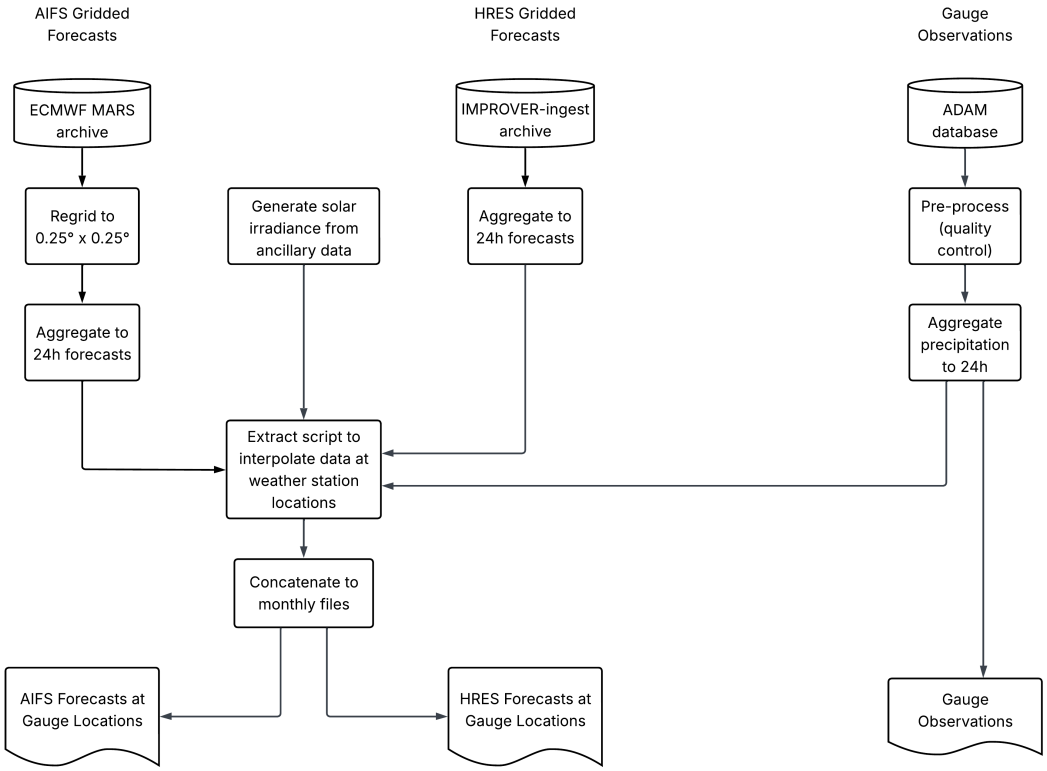

Figure 1: An overview of the pre-processing steps required to create the training data in this study.

2.1.1 AIFS pre-processing

AIFS data was downloaded from the ECMWF MARS archive using the ECMWF Web API (ECMWF, 2019).

The AIFS forecast data correspond to a N320 reduced Gaussian grid (ECMWF, 2023a) which differs from the regular latitude-longitude grid typically used in the Australian domain (Glowacki et al., 2012). The AIFS data were re-gridded using the `earthkit-regrid` library (ECMWF, 2023b) to a $0.25° \times 0.25°$ grid. Like HRES, AIFS is a global model, so the domain was restricted to latitudes between -60 and 5 and longitudes between 75 and 180, matching the outputs of the Bureau's operational data ingest system for HRES. The study domain is shown in Figure 2.

AIFS forecasts were extracted from the MARS archive at six-hourly temporal resolution. For this study, forecasts from the 12 UTC base time were aggregated to 24-hour periods to produce daily forecasts valid at 12 UTC, with lead-times ranging from 24 to 240 hours. Precipitation fields were provided as accumulations from the forecast reference time; 24-

hour precipitation totals were therefore calculated by differencing successive accumulation fields. Wind speed was aggregated as the mean over the corresponding 24-hour period. AIFS provides wind as two perpendicular vector components, which were combined using the Euclidean norm to produce a scalar wind speed magnitude. Directional information was not retained, as RainForests does not explicitly use spatial or directional based features.

2.1.2 HRES pre-processing

HRES data was obtained from the Bureau's IMPROVER-ingest archive, which contains NWP data that has been pre-processed for input into IMPROVER. This processing includes converting to NetCDF format, restricting to the Australian domain, and calculating the wind magnitudes. The only remaining processing required was to aggregate the forecasts into 24-hour periods.

2.1.3 Observational data

Gauge observations are used for model training and quantitative verification. Observations from automatic weather stations (AWS) were extracted from the Bureau's ADAM observation database. Data was filtered to include only rows with quality status "Accepted", and times having fewer than 50 observations in the one-hour period were excluded. After filtering, there were 777 weather stations represented in the data, as shown in Figure 2. Hourly totals were obtained by aggregating data from one-minute observations.

Satellite data from the Global Precipitation Measurement (GPM) mission satellite (Hou et al., 2014) is used for qualitative verification.

2.1.4 Clear sky solar radiation data

Solar radiation data were derived using the empirical model of Ineichen and Perez (2002), rather than being directly provided by the HRES or AIFS forecasts. An existing IMPROVER function implementing this model was used, which generates clear-sky solar radiation from surface altitude, location, and time of year.

2.2 Site extraction

To create training data for the RainForests model, we extracted the training features at AWS locations across Australia. For each site, we extracted the forecast data at AWS sites from the surrounding grid points. For the precipitation and CAPE variables, we used nearest-neighbour interpolation in order to preserve zeros and very high values, while for wind speed and solar radiation we used linear interpolation.

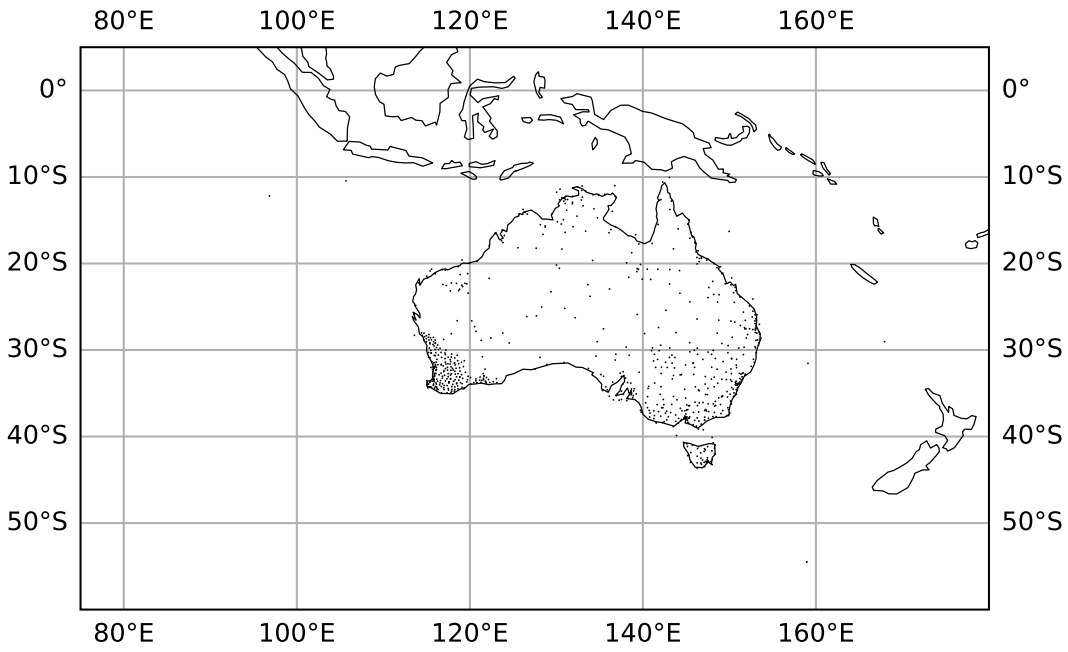

Figure 2: Locations of automatic weather stations used for gauge observations, plotted on the study domain of latitude $[-60, 5]$ and longitude $[75, 180]$.

2.3 Models

Once the training dataset had been created, we then replicated the methods described in the original RainForests paper for creating a post-processing model (Trotta et al., 2024b). The training dataset consisted of seven non-consecutive months of forecasts covering all seasons to provide a broad range of weather contexts (see Table 1). To simplify data handling, forecasts were grouped by month according to forecast reference time. The validation set was used for parameter optimisation of the Rainforests models, and the test set was used for evaluating results.

Quality control was applied to the observational rainfall data prior to training. Gauge observations were excluded if the minimum reported measurement exceeded 0.2 mm, if measurements exceeded known Australian record rainfall amounts or if precipitation values were likely attributable to dew rather than rainfall. For more detail on the quality control methods see Trotta et al. (2024a, p. 4). After filtering, 762 stations were used in the training data.

Train	Validation	Test
03–2024	06–2024	07–2024
04–2024	09–2024	10–2024
05–2024	12–2024	01–2025
08–2024	03–2025	04–2025
11–2024	06–2025	07–2025
02–2025		
05–2025		

Table 1: Training, validation, and test period splits.

2.3.1 Training Features

The RainForests models were trained using the following predictor variables:

- total precipitation (m)
- convective precipitation (m)
- wind speed ($\mathrm{m/s}$, at 700 hPa for AIFS and 800 hPa for HRES)[1]
- accumulated clear-sky solar radiation ($\mathrm{Ws/m^2}$)
- CAPE ($\mathrm{J/kg}$, HRES only)[2]

For wind speed and CAPE, the model inputs were aggregations of the forecasts at 6 h lead-time frequency (that is, the value for the 24 h lead-time is the aggregation of the values at lead-times 6, 12, 18, and 24). These four forecasts were aggregated by taking the mean for wind speed, and the maximum for CAPE. The extracted datasets were then aggregated into monthly files to be split into training, validation and testing datasets.

2.3.2 Model Structure

In the original RainForests methodology, a separate gradient-boosted decision tree (GBDT) is trained for each combination of forecast lead-time and rainfall threshold. Each model predicts whether accumulated precipitation exceeds a specific threshold for a specific lead-time. We used the same set of thresholds as the Bureau's operational implementation of

[1]The 700 hPa wind speed is considered to have the best predictive power, but is not available in BoM's IMPROVER ingest archive of HRES data.

[2]CAPE is only included for HRES, this variable is not available in AIFS. As of November 2024, the ECMWF NWP forecast no longer includes the CAPE variable; see Roberts et al., 2024. At around that time, the Bureau's operational data ingest workflow transitioned to using maximum uncertainty CAPE (MUCAPE) instead. Previous work by Benjamin Owen (not published) investigated applying existing operational RainForests models, trained on CAPE, to ECMWF NWP forecasts containing either CAPE or MUCAPE and found that these produce very similar outputs.

Rainforests. These thresholds are more tightly concentrated at lower rainfall values with larger gaps at higher values. Specifically, the 26 thresholds used are (mm): 0.05, 0.1, 0.2, 0.4, 0.6, 1, 2, 5, 7, 10, 15, 25, 35, 50, 75, 100, 125, 150, 200, 250, 300, 350, 400, 450, and 500.

In this study, we also evaluate an alternative training strategy. Instead of training separate models for each lead-time, lead-time is included as an input feature. A single GBDT is then trained for each rainfall threshold and applied across all lead-times. This combined approach increases the amount of training data available to each model by a factor of ten. However, it may reduce the model's ability to make lead-time-specific corrections, particularly at the shortest and longest lead-times.

Under the original approach, 26 rainfall thresholds and 10 lead-times result in 260 separate GBDT models. Under the combined approach, only 26 GBDT models are required. Each GBDT then produces an estimated probability that accumulated rainfall exceeds its corresponding threshold given forecasted conditions and lead-time.

Validation and test forecasts were calibrated using the original RainForests calibration function, with our newly trained GBDT models applied via the IMPROVER Command Line Interface (CLI). We compared both the original 260 model approach and the updated combined-lead-time approach with 26 models. The calibration step combines the discrete threshold exceedance probabilities into a continuous cumulative distribution function (CDF) for rainfall by interpolating between thresholds and enforcing non-decreasing probabilities with increasing thresholds.

The calibrated output is produced as a NetCDF cube representing a probabilistic forecast where a probability is assigned to each rainfall threshold for a specific lead-time and location. This same function can be applied to our gridded datasets created earlier, not just the training data at gauge sites.

2.3.3 Parameter optimisation

The Optuna package (Akiba et al., 2019) was used to systematically tune model parameters. A single execution of the training script could take up to one hour, therefore parameter optimisation was performed on a small subset of the data. This subset was constructed by selecting representative lead-times and thresholds. Specifically, the following values were used:

- Thresholds (mm): $\{0.1, 0.2, 0.4, 1, 10, 50, 100\}$
- lead-times (days): $\{1, 5, 10\}$

The optimisation proceeded by varying parameters systematically according to a loss function evaluated on the corresponding validation set. Two optimisation strategies were consid-

ered. Firstly, the Continuous Ranked Probability Score (CRPS) (Hersbach, 2000) was used as a combined loss function to find a single set of parameters for all rainfall thresholds. For the second approach we sought to find different sets of parameters at each representative threshold, with the Brier score (Brier, 1950) used as the loss function.

Initially we chose to optimise several parameters that controlled the learning rate, tree structure and regularisation:

1. Learning rate
2. Number of trees
3. Number of leaves
4. L1 regularisation
5. L2 regularisation
6. Minimum data required to make a leaf

This proved to be too many parameters for the optimisation algorithm to effectively converge on a solution. In a second iteration we focused simply on optimising tree size, number of leaves and L1 regularisation. The threshold-specific parameter approach failed to converge and the combined parameter result suggested a moderate increase of leaves per GBDT and fewer trees overall. However, preliminary results indicated that the GBDT models were relatively insensitive to this parameter choice. Additionally, since only a small subset of the data could be used for parameter optimisation, the selected parameter values failed to meaningfully translate to improvements when applied to the full dataset.

Since the parameter optimisation failed to yield any significant improvement, we decided to proceed with the current operational RainForests setup, where each GBDT model was configured with a maximum of 400 decision trees. Individual trees were deliberately shallow, with a maximum of five leaves, ensuring that each model remained simple and robust while allowing the ensemble to capture non-linear relationships between predictors and rainfall occurrence.

3 Results and Discussion

The central objective of this study is to determine whether RainForests calibration is effective when paired with AIFS forecasts; we benchmarked this against HRES forecasts undergoing the same calibration process. Overall, calibrated AIFS shows consistently better forecast skill than calibrated HRES across the majority of forecast lead-times and thresholds, as demonstrated by the validation metrics presented in Section 3.1. This is consistent with the performance of uncalibrated AIFS, which Lang et al. (2024) found to match or exceed the accuracy of uncalibrated HRES. Notably, AIFS achieves this greater accuracy while operating at a coarser spatial resolution of $0.25°$ versus $0.1°$.

We will begin by comparing the quantitative performance of the calibrated models, then proceed to a case study to compare their qualitative characteristics.

3.1 Quantitative evaluation of post-processed forecasts

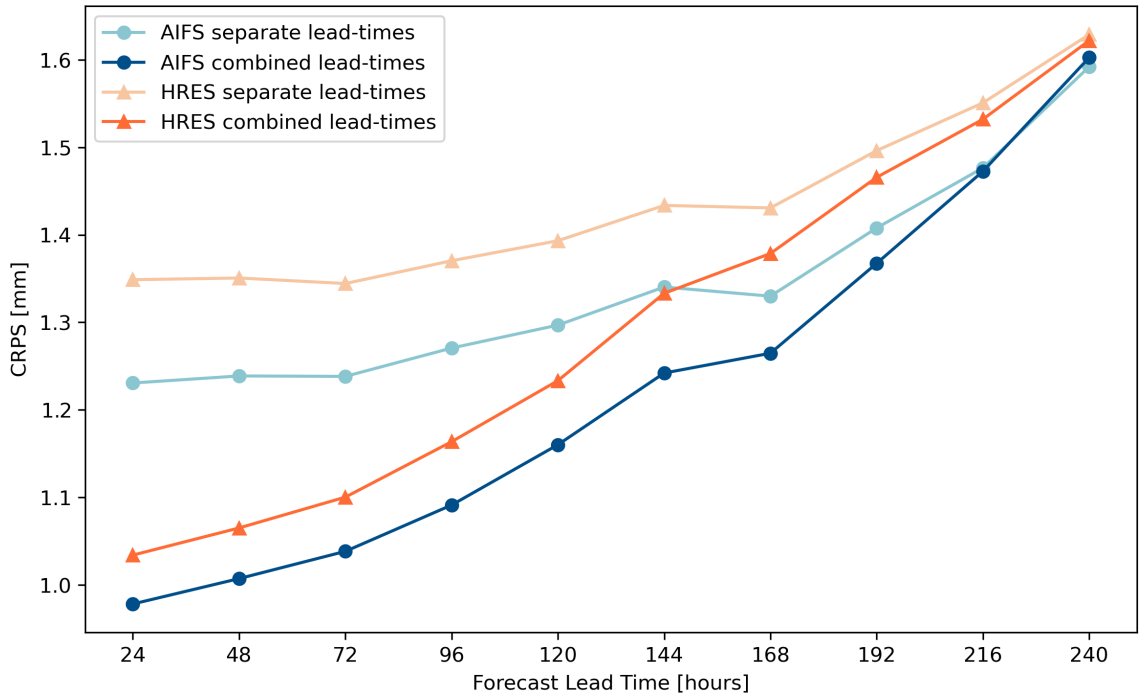

Figure 3: Continuous Ranked Probability Score (CRPS) for the original and combined lead-time post-processing approaches, evaluated across forecast lead-times from 24 to 240 hours.

We use the scores package (Leeuwenburg et al., 2024) for calculating quantitative metrics.

The validation metric used for probabilistic forecasts is typically the continuous ranked proba-

bility score (CRPS) (see details at (Scores 2.3.0 documentation, 2026b)). This metric represents the average discrepancy between the forecast CDF and the CDF of the corresponding perfect forecast. A smaller CRPS value is better. As illustrated in Figure 3, AIFS consistently outperforms HRES across all lead-times in terms of CRPS, though the performance gap narrows at longer lead-times. This behaviour is consistent with the superior deterministic accuracy of uncalibrated AIFS, which exhibits approximately 9.5% lower RMSE than uncalibrated HRES when averaged across lead-times (Figure 4). Together, these results indicate that RainForests calibration preserves the deterministic skill advantage of AIFS while successfully transforming its output into a probabilistic forecast.

Figure 3 also compares the original lead-time-specific RainForests approach with the combined approach, in which a single model is trained per rainfall threshold. The combined approach yields substantially lower CRPS at shorter lead-times, with a larger improvement for HRES than for AIFS. The improvement likely stems from the relatively small training dataset available for this study. It is likely that the original training approach suffers from data sparsity and the combined approach allows each GBDT to learn from ten times more data. This increase in training data appears to outweigh the loss of lead-time-specificity, as the combined approach achieves lower CRPS at all forecast lead-times except 240 hours.

As a consequence of these improved results, all subsequent results utilise the combined-lead-time approach.

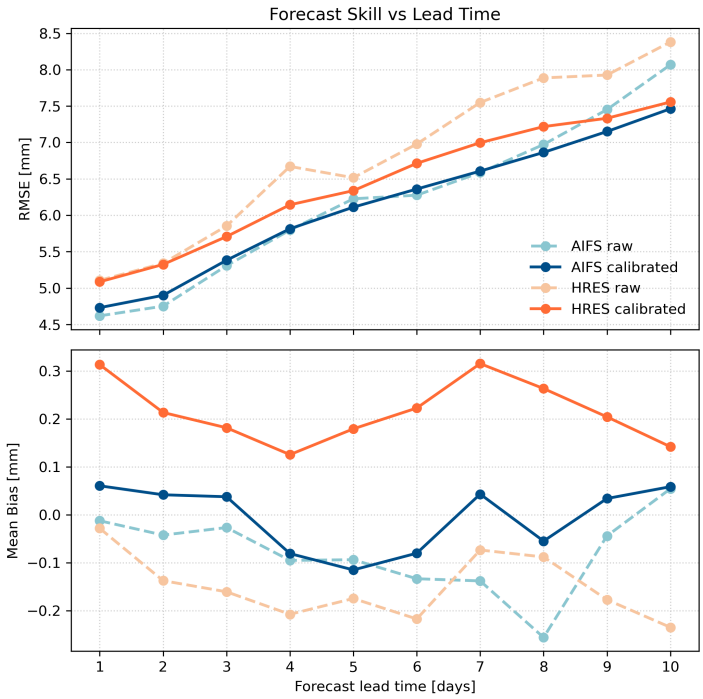

Figure 4: RMSE and mean bias (forecast minus observations) calculated from expected value of probabilistic forecasts. For these metrics we compare RainForests calibrated models to uncalibrated (raw) models at all forecast lead-times.

To compare performance of the model before and after calibration we first derive an expected value from the probabilistic forecast. This is done by integrating the piecewise-linear CDFs of the probabilistic forecasts. The expected value allows us to directly compare the calibrated forecast with the deterministic values of the raw forecast. In Figure 4 we can see the forecast error (RMSE) and the mean bias which allows us to determine the "direction" of the error. HRES benefits more significantly from calibration in terms of RMSE, however AIFS remains more accurate at all lead-times. The fact that post-processing yields only limited improvement in RMSE for AIFS is expected, as the model is already optimised for this metric during training. However, some gains are observed at lead-times of 8-10 days. Calibrated HRES displays a clear tendency to overestimate forecasts: mean-bias across all lead-times is 0.217 mm compared to -0.005 mm for calibrated AIFS. For both AIFS and HRES, calibration tends to increase the forecast values overall. This is likely related to the tendency of the calibrated models to consistently estimate small rather than zero rainfall amounts as can be observed in Figures 7 and 8. While calibration has little effect on the RMSE of AIFS, it provides substantial value when quantified by CRPS, a more critical metric for assessing the quality of a probabilistic forecast. Our post-processing approach is likely not optimal for producing an expected-value forecast, since it models the probability distribution as piecewise-linear which is only an approximation of the true distribution. It is likely

that a more accurate expected-value forecast could be obtained by optimising a calibration function for the expected value directly.

Table 2: Brier score at predicted rainfall thresholds for calibrated forecasts at a 24 h lead-time. A lower score is better, AIFS exhibits a lower Brier score across most thresholds.

Rainfall Threshold (mm)	AIFS with RainForests	HRES with RainForests
0.00	0.00	0.00
0.01	1.27×10^{-1}	1.27×10^{-1}
0.05	1.27×10^{-1}	1.27×10^{-1}
0.10	1.27×10^{-1}	1.27×10^{-1}
0.20	1.13×10^{-1}	1.15×10^{-1}
0.40	8.19×10^{-2}	8.55×10^{-2}
0.60	7.42×10^{-2}	7.82×10^{-2}
1.00	6.86×10^{-2}	7.25×10^{-2}
2.00	5.90×10^{-2}	6.28×10^{-2}
5.00	4.42×10^{-2}	4.69×10^{-2}
7.00	3.75×10^{-2}	3.96×10^{-2}
10.00	3.05×10^{-2}	3.20×10^{-2}
15.00	2.17×10^{-2}	2.28×10^{-2}
25.00	1.08×10^{-2}	1.14×10^{-2}
35.00	5.70×10^{-3}	5.98×10^{-3}
50.00	2.39×10^{-3}	2.53×10^{-3}
75.00	8.26×10^{-4}	9.14×10^{-4}
100.00	4.93×10^{-4}	5.22×10^{-4}
125.00	2.46×10^{-4}	2.65×10^{-4}
150.00	1.67×10^{-4}	1.72×10^{-4}
200.00	6.08×10^{-5}	6.72×10^{-5}
250.00	1.22×10^{-5}	1.92×10^{-5}
300.00	7.27×10^{-6}	1.18×10^{-5}
350.00	2.31×10^{-6}	6.02×10^{-7}
400.00	7.23×10^{-7}	5.85×10^{-7}
450.00	2.26×10^{-8}	1.74×10^{-7}
500.00	2.26×10^{-8}	1.74×10^{-7}

The Brier score is a common metric used to evaluate binary probability forecasts; it is very similar to MSE. A smaller Brier score is better where a perfect score is 0 and the worst score is 1 (Scores 2.3.0 documentation, 2026a). In Table 2 we have calculated the Brier score for both models at the 24 h lead-time. We can see that AIFS performs better at most thresholds. At very high thresholds, data sparsity leads to increased noise in the results.

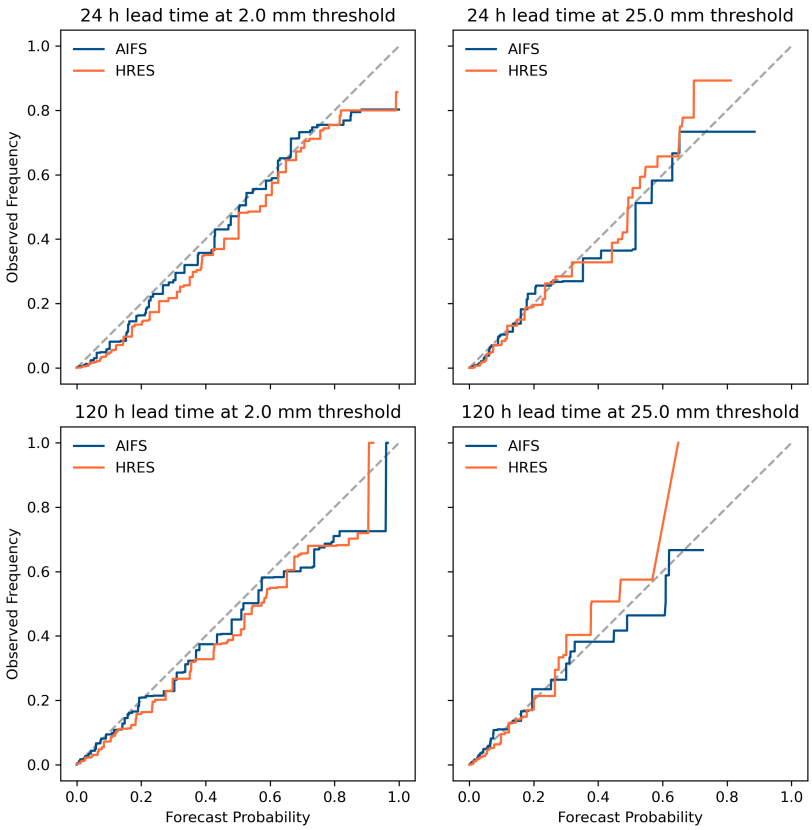

Figure 5: Reliability diagram for AIFS and HRES at 24 h and 120 h lead-times. Calculated using isotonic regression. Values closer to the diagonal line indicate better calibration. Missing values on the forecast probability axis indicate the model never made a prediction with that level of certainty.

Reliability curves provide a visual assessment of how well probabilistic forecasts are calibrated across different probability thresholds. We use isotonic regression as implemented in the Scores package (Scores 2.3.0 documentation, 2026c), to visualise the relationship between model confidence and observed probabilities for a given rainfall threshold. Isotonic regression is a similar process to linear regression but attempts to model a non-decreasing relationship. In Figure 5 we can see AIFS and HRES forecasts at the 24 h lead-time for two different rainfall thresholds. Both models perform similarly at the 2.0 mm threshold, however HRES demonstrates a slight tendency towards overconfidence. At the 25.0 mm threshold both models are well calibrated at low probabilities, but at higher probabilities the data is sparse and it is difficult to draw definite conclusions. For longer lead-times such as 120 h, the reliability of both models diminishes to a similar extent.

3.2 Qualitative evaluation of post-processed forecasts

It is very difficult for models to accurately predict probabilities for large rainfall thresholds because high rainfall amounts are uncommon, so there is little training data for the models to learn from. As a result, when a forecast model confidently predicts a large amount of rainfall, the post-processing model will often temper this prediction to a smaller amount closer to the mean. This is problematic because very high rainfall values are often the most consequential. Ideally, our model will confidently predict high rainfall values without sacrificing reliability.

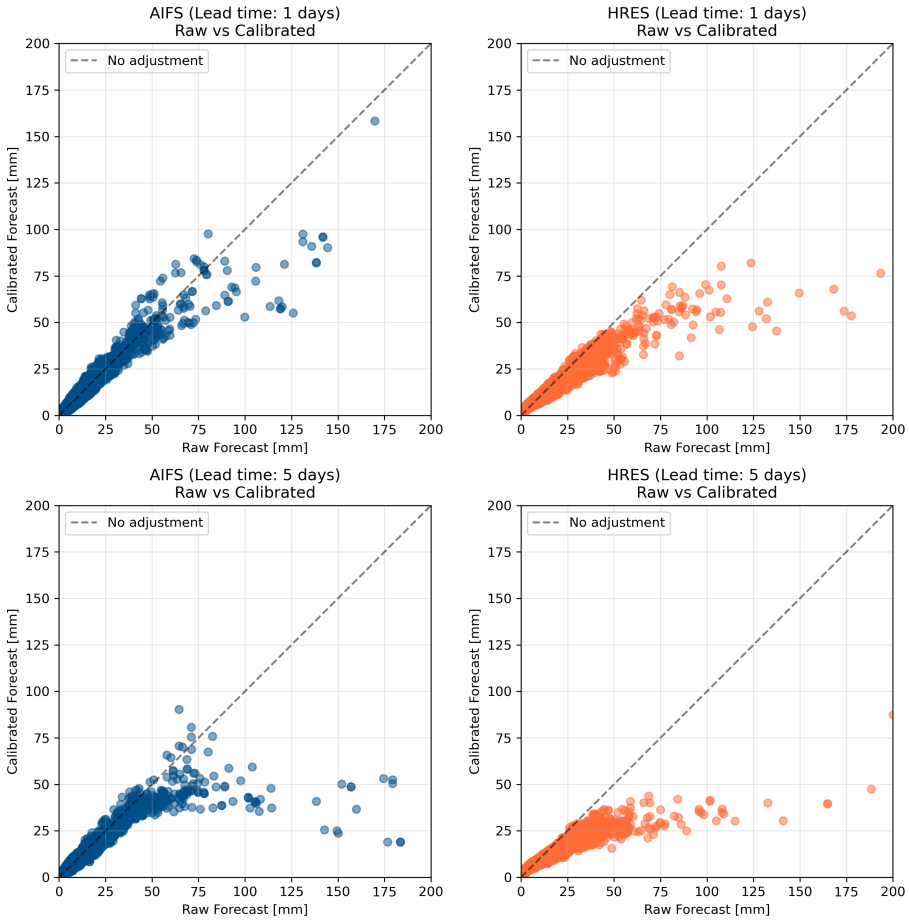

Figure 6: Scatter plot comparing raw forecast values on the horizontal axis and expected value of calibrated forecasts on the vertical axis. Points below the dotted line indicate that the predicted rainfall amount is reduced by RainForests calibration.

In Figure 6 the dotted line represents the situation where the expected value of the calibrated forecast is equal to the raw forecast. Points below the line occur when the calibration reduces the predicted rainfall amount. Both the AIFS and HRES charts represent the same set of

observations. The HRES forecasts are being suppressed more than those of AIFS by the calibration procedure. This difference likely reflects the closer agreement between raw AIFS forecasts and the observation data, which reduces the need for the post-processing model to suppress larger forecasted rainfall amounts.

If we increase the lead-time this pattern remains largely the same for both AIFS and HRES except that both raw forecasts temper their higher rainfall amounts. At longer lead-times, RainForests tends to reduce all HRES forecasts exceeding approximately 20 mm. Considering these results alongside Figure 4 – which shows that RainForests calibration overall *increases* the raw forecast, resulting in a positive mean bias – suggests that the positive bias is driven by RainForests consistently increasing low rainfall amounts in HRES forecasts.

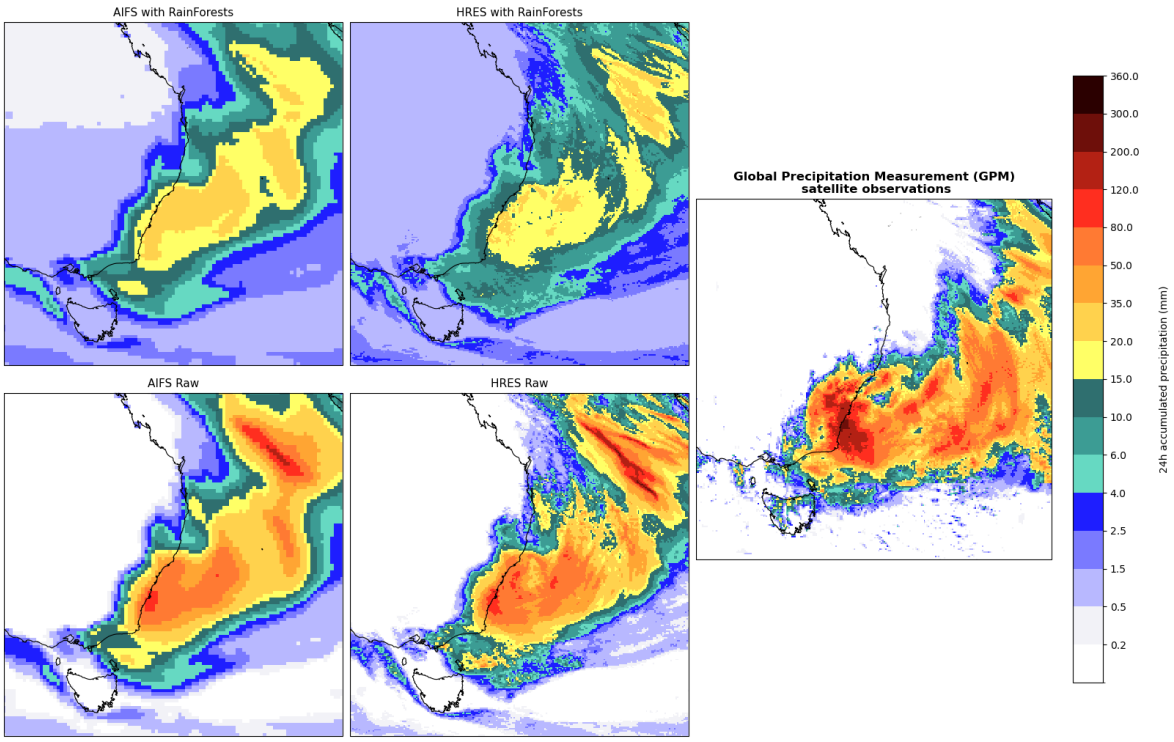

Figure 7: Case study of an extreme rainfall event on the New South Wales coast with over 200 mm in 24 hours in some locations. Forecasts at 24-hour lead-time valid at 2025-07-02 1200 UTC. From left to right: AIFS gridded forecasts, HRES forecasts, and GPM satellite observations. Top panels show calibrated forecasts, and bottom panels show raw forecasts. Calibrated forecasts exhibit widespread overestimation of light rainfall. Note the coarser resolution of AIFS compared to HRES.

From the 1st to the 3rd of July 2025 extreme rainfall fell along the NSW coast. Some areas received over 200 mm of rainfall in a 24 h period in this time. This event offers a good case study for how these calibrated models perform in cases of high rainfall. In Figure 7 we can see that for both the calibrated HRES and AIFS forecasts, the predicted rainfall was drastically reduced compared to the raw forecasts. When we compare this to the Global

Precipitation Measurement (GPM) satellite observations (Hou et al., 2014) the raw forecasts more closely fit the shape and intensity of the rainfall. This is one of the highest rainfall days across the test or training data so there would be limited examples for the training models to improve accuracy for such events. Nonetheless, we can see that while both raw forecasts look very similar, the calibrated AIFS forecast retains a higher rainfall forecast compared to HRES which in this case is more accurate. We can again see examples of the positive bias described by Figure 4, where areas with zero rainfall in the observations have small non-zero amounts in the forecast. This overestimation of light precipitation could be a side effect of calculating the expected value from a probabilistic forecast, or could be due to a lack of model confidence resulting from insufficient training data. When producing the same image using the original training method with one tenth of the training data (using the separate lead-time models) this effect was even more pronounced, suggesting that insufficient training data may be the cause. The most recent AIFS version, introduced after the time period used for this study, has improved forecasting of light precipitation amounts (Moldovan et al., 2025), so this issue may be mitigated by using the newer version. In Figure 8, a 24 h forecast across Australia is presented for a randomly selected day in the test dataset. This again presents the light precipitation issue, however it is less pronounced for the calibrated AIFS model.

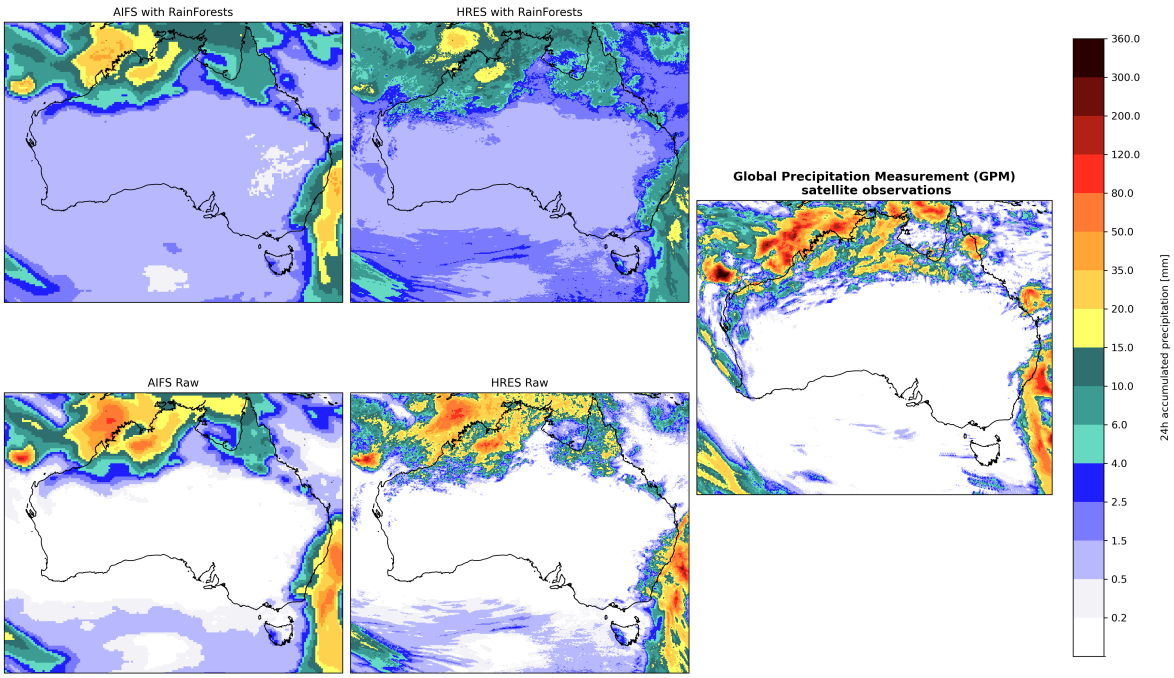

Figure 8: Forecasts at 24 h lead-time with valid time 2025-01-17 1200 UTC. From left to right: AIFS gridded forecasts, HRES forecasts, and GPM satellite observations. Top panels show calibrated forecasts, and bottom panels show raw forecasts. Calibrated HRES appears to under-forecast high values of rainfall and over-forecast light rainfall.

4 Conclusion

The Bureau's current operational precipitation forecast is produced using a weighted blend of physics-based numerical weather prediction models (Trotta et al., 2024a). Recent advances in ML-based global forecasting systems, most notably ECMWF's Artificial Intelligence Forecasting System (AIFS) (Lang et al., 2024), present an opportunity to enhance this framework by incorporating the complementary strengths of ML-based forecasts.

The results of this study demonstrate that AIFS performs favourably within this rainfall post-processing context. In particular, the interaction between AIFS and RainForests produces high-quality probabilistic rainfall forecasts and consistently outperforms HRES when both models are subjected to the same post-processing methodology.

Although the expected value of the calibrated forecast struggled with over-prediction of light rainfall, overall performance as measured by Brier Score and Continuous Ranked Probability Score (CRPS) was encouraging. Given that RainForests has previously been validated for physics-based models similar to HRES (Trotta et al., 2024a), the superior performance of AIFS under the same processing framework provides evidence that AIFS could perform well if introduced into the Bureau's operational precipitation forecast blend. As AIFS continues to develop, with a substantial update to rainfall forecasting capabilities released in August 2025 (Moldovan et al., 2025), its potential value in an operational setting is likely to increase further.

A key limitation of this study is the relatively small training dataset, reflecting the fact that AIFS is still a comparatively new forecasting system. Substantial benefits were observed when using a combined model for different lead-times. However, this improvement may largely be due to the increased training dataset per model compared to the original approach. When more training data are available it would be valuable to compare the combined and original training approaches.

Parameter optimisation appeared to have little effect on model accuracy but this may have been due to practical limitations in searching the full parameter space. Future forecast improvements may be found in the exploration of alternative model architectures, such as fully connected neural networks, which can offer greater flexibility in representing complex relationships in the data.

Overall, the results of this study support further investigation into the operational use of AIFS for rainfall forecasting at the Bureau, particularly in combination with existing ML-based post-processing systems such as RainForests.

Acknowledgements

This work was completed as part of Felix Esperson's Masters of Data Science at La Trobe University. We thank the university for facilitating this collaboration. We thank Benjamin Owen and Robert Johnson for helpful feedback on a draft of this report. Our work builds on code developed by current and past members of the Forecast Improvement team, and we are grateful for their contributions. We thank the European Centre for Medium-Range Weather Forecasts (ECMWF) for making available data from the HRES and AIFS forecasts, and the National Computing Infrastructure (NCI) for providing computing facilities.

References

Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama (2019). "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Brier, G. W. (1950). "Verification of forecasts expressed in terms of probability". In: *Monthly Weather Review* 78.1, pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

de Burgh-Day, C. O. and T. Leeuwenburg (2023). "Machine Learning for Numerical Weather and Climate Modelling: A Review". In: *Geoscientific Model Development* 16.22, pp. 6433–6477. DOI: 10.5194/gmd-16-6433-2023.

ECMWF (2019). *ecmwf-api-client*. URL: https://github.com/ecmwf/ecmwf-api-client.

– (2023a). *AIFS Machine Learning Data*. URL: https://www.ecmwf.int/en/forecasts/dataset/aifs-machine-learning-data.

– (2023b). *earthkit-regrid*. URL: https://github.com/ecmwf/earthkit-regrid.

Glowacki, T. J., Y. Xiao, and P. Steinle (2012). "Mesoscale Surface Analysis System for the Australian Domain: Design Issues, Development Status, and System Validation". In: *Weather and Forecasting* 27.1, pp. 141–157. DOI: 10.1175/waf-d-10-05063.1.

Hersbach, H. (2000). "Decomposition of the Continuous Ranked Probability Score for ensemble prediction systems". In: *Weather and Forecasting* 15.5, pp. 559–570. DOI: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hou, A. Y., R. K. Kakar, S. Neeck, A. A. Azarbarzin, C. D. Kummerow, M. Kojima, R. Oki, K. Nakamura, and T. Iguchi (2014). "The Global Precipitation Measurement Mission". In: *Bulletin of the American Meteorological Society* 95.5, pp. 701–722. DOI: 10.1175/BAMS-D-13-00164.1.

Ineichen, P. and R. Perez (2002). "A New Airmass Independent Formulation for the Linke Turbidity Coefficient". In: *Solar Energy* 73.3, pp. 151–157. DOI: 10.1016/s0038-092x(02)00045-2.

Lang, S., M. Alexe, M. Chantry, J. Dramsch, F. Pinault, B. Raoult, M. C. A. Clare, C. Lessig, M. Maier-Gerber, L. Magnusson, Z. B. Bouallègue, A. P. Nemesio, P. D. Dueben, A. Brown, F. Pappenberger, and F. Rabier (2024). *AIFS – ECMWF's Data-Driven Forecasting System*. arXiv: 2406.01465 [physics.ao-ph].

Leeuwenburg, T., N. Loveday, E. E. Ebert, H. Cook, M. Khanarmuei, R. J. Taggart, N. Ramanathan, M. Carroll, S. Chong, A. Griffiths, and J. Sharples (2024). "scores: A Python package for verifying and evaluating models and predictions with xarray". In: *Journal of Open Source Software* 9.99, p. 6889. DOI: 10.21105/joss.06889.

Moldovan, G. et al. (2025). “AIFS 1.1.0: An update to ECMWF’s machine-learned weather forecast model AIFS”. In: *EGUsphere* 2025, pp. 1–23. DOI: 10.5194/egusphere-2025-4716.

Roberts, C. D., B. Ingleby, A. Geer, E. Hólm, M. Janousek, F. Prates, and M. Rodwell (2024). *IFS upgrade improves near-surface wind and temperature forecasts*. URL: https://www.ecmwf.int/en/newsletter/181/earth-system-science/ifs-upgrade-improves-near-surface-wind-and-temperature.

Roberts, N. et al. (2023). “IMPROVER: The New Probabilistic Postprocessing System at the Met Office”. In: *Bulletin of the American Meteorological Society* 104.3, E680–E697. DOI: 10.1175/BAMS-D-21-0273.1.

Scores 2.3.0 documentation (2026a). *Brier Score — Scores 2.3.0 Documentation*. URL: https://scores.readthedocs.io/en/stable/tutorials/Brier%5C_Score.html.

– (2026b). *Continuous Ranked Probability Score (CRPS) — Scores 2.3.0 Documentation*. URL: https://scores.readthedocs.io/en/stable/tutorials/CRPS%5C_for%5C_CDFs.html.

– (2026c). *Isotonic Regression and Reliability Diagrams*. URL: https://scores.readthedocs.io/en/stable/tutorials/Isotonic%5C_Regression%5C_And%5C_Reliability%5C_Diagrams.html.

Trotta, B., J. Canvin, T. Gale, T. Hume, R. Johnson, J. Liu, D. Mentiplay, B. Owen, A. Schubert, G. Weymouth, and J. Whelan (2024a). “An Initial Benchmarking of IMPROVER – Part 2: Evaluation of Precipitation Diagnostics”. In: *Bureau of Meteorology Research Reports* 93.

Trotta, B., R. Johnson, C. de Burgh-Day, D. Hudson, E. Abellan, J. Canvin, A. Kelly, D. Mentiplay, B. Owen, and J. Whelan (2025). “Statistical Postprocessing Yields Accurate Probabilistic Forecasts from Artificial Intelligence Weather Models”. In: *Artificial Intelligence for the Earth Systems* 4.4. DOI: 10.1175/aies-d-25-0037.1.

Trotta, B., B. Owen, J. Liu, G. Weymouth, T. Gale, T. Hume, A. Schubert, J. Canvin, D. Mentiplay, J. Whelan, and R. Johnson (2024b). “RainForests: A Machine Learning Approach to Calibrating NWP Precipitation Forecasts”. In: *Weather and Forecasting* 39.11, pp. 1715–1732. DOI: 10.1175/WAF-D-23-0211.1.

Van Poecke, A., L. Meuris, M. Cisneros, M. Van Ginderachter, P. Hellinckx, and H. Tabari (2025). “A Comparative Sensitivity Analysis of Loss Functions in Machine Learning-Based Weather Forecasting”. In: *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*. Ed. by L. Barolli. Cham: Springer Nature Switzerland, pp. 318–326.

Vannitsem, S. et al. (2021). “Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World”. In: *Bulletin of the American Meteorological Society* 102.3, E681–E699. DOI: 10.1175/BAMS-D-19-0308.1.