

Article

Comparison of Physical and Deep Learning Weather Forecast Models for Agricultural Decision-Making in Belgium

Valérien Authelet, Sébastien Dandrifosse *, Valéry Michaud, Jean Pierre Huart, Viviane Planchon 
and Damien Rosillon

Walloon Agricultural Research Centre, 9 rue de Liroux, 5030 Gembloux, Belgium;
v.planchon@cra.wallonie.be (V.P.); d.rosillon@cra.wallonie.be (D.R.)

* Correspondence: s.dandrifosse@cra.wallonie.be

Abstract

Weather forecasts are crucial in agricultural decision-making. The first objective was to assess the performance of eight weather forecast models for predicting five key meteorological variables in agriculture: air temperature, humidity, wind speed, global radiation, and precipitation. The tested models were ICON-D2, AROME, HARMONIE, MAR, GFS, and three models from the ECMWF: the deterministic model (HRES), the ensemble model (ENS), and the deep learning-based model (AIFS). The forecasts were compared over a six-month period with observations from 22 weather stations located in Wallonia (Belgium). AIFS achieved the lowest RMSE for predicting global radiation. For lead times up to 48 h, ICON-D2 had the lowest RMSE for wind speed; ICON-D2 and AIFS showed the lowest RMSE for air temperature and humidity, and AIFS and HRES performed best at predicting rainy versus non-rainy days. For lead times beyond 48 h, AIFS had the lowest RMSE for predicting air temperature and humidity, and ENS had the lowest RMSE for predicting wind speed. The second objective was to evaluate the suitability of these forecasts to feed a simple agricultural decision-support tool, helping farmers identify the optimal time windows for spraying plant protection products. ICON-D2 and AIFS yielded the most reliable recommendations.

Keywords: forecast; verification; agriculture; spraying; decision support tool; AI; precipitation; temperature; wind speed; global radiation

1. Introduction

Weather forecasts are crucial in agricultural decision-making, as weather is one of the biggest concerns for farmers. Some farmers check weather forecasts several times a day to plan activities such as planting, crop protection, fertilisation, irrigation, and harvesting.

The performance of forecasts within a 3- to 10-day horizon has improved by about one day per decade. This means that current 6-day forecasts are as accurate as 5-day forecasts were 10 years ago [1]. However, despite these improvements, weather forecasts are still subject to many uncertainties. Farmers and agronomists have access to several forecast sources and wonder which forecasts to trust. Hence, there is a need to compare and evaluate weather forecast models.

A few studies have evaluated physical forecast models for predicting near-surface weather variables relevant to agriculture, i.e., air temperature at 2 m, relative humidity, wind speed at 2 m, precipitation and global radiation. The European Centre for Medium-Range Weather Forecasts (ECMWF) regularly provided reports on the evaluation of its models, including performance scores for near-surface variables [2]. Green Books are annual reports produced by ECMWF members and give details on how national meteorological services



Academic Editor: Gianni Bellocchi

Received: 22 May 2026

Revised: 22 July 2026

Accepted: 24 July 2026

Published: 26 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

use, verify and validate the forecasts of the ECMWF in their country. At the time of writing, the most recent Green Book was published in 2024 (Green Book chapters are available on the ECMWF website, consulted on 19 February 2026). The last report for Belgium was published in 2021. It evaluated the performance of ECMWF deterministic and ensemble models for one year. Key variables studied included 2-metre air temperature, 10-metre wind speed, and precipitation accumulated over 6 h. The U.S. National Oceanic and Atmospheric Administration (NOAA) published verification results of its model on its website, comparing it with models from the ECMWF and the national weather services of the UK, Germany, and France (noaa.gov consulted on 24 February 2026). Meteoblue (meteoblue AG, Basel, Switzerland), a meteorological services company, also published reports on the performance of forecast models for several near-surface weather variables.

In the peer-reviewed literature, Haiden et al. [3] compared precipitation from several forecast models with ground-based observations collected worldwide. Magnaldo et al. [4] evaluated solar radiation forecasts from the AROME model in detail and proposed improvements to enhance its performance. Hagelin et al. [5] tested the AROME model for nowcasting air temperature and wind speed in France. In addition, several forecast models have been evaluated in student theses. Häuser [6] analysed the performance of four global and two regional models for wind speed prediction. Reiß [7] assessed solar radiation forecasts from several models on a global scale and proposed a multi-model approach to reduce prediction errors. Fessler [8] studied the air temperature forecasts of six global models.

Most of these studies focused on a limited number of models or a restricted set of meteorological variables. Moreover, weather forecast models evolve rapidly. Regular updates and improvements can substantially alter their performance, meaning that earlier evaluations may quickly become obsolete. At the same time, physics-based forecast models are being reshaped by advances in artificial intelligence (AI), particularly by machine learning and deep learning techniques. These approaches have already demonstrated considerable potential [9].

Waqas et al. [10] conducted a literature review on the integration of machine learning and deep learning into numerical weather prediction models. These models make predictions very quickly and require less computational energy than conventional models. Their performance is largely determined by the number and quality of training data. However, many challenges still need to be addressed to improve these data-driven approaches [11]. AI models may struggle to predict extreme weather events, especially because such atmospheric conditions are poorly represented in the input data [12]. They also lack interpretability [10]. ECMWF developed its own AI model, named AIFS. The deterministic model has been operational since February 2025, and the ensemble model has been operational since July 2025. AIFS produced very accurate forecasts for upper-level variables and surface variables [13]. Other AI models such as Aurora [14], Pangu-Weather [15], and GraphCast [16] also gave promising results. Yet these models are new and require further evaluation and comparison with conventional physics-based models.

This study addresses this gap by comparing the performance of eight recent weather forecast models, including a deep learning-based model, for predicting five key meteorological variables relevant to agricultural decision-making in Belgium. The objectives were (i) to assess forecasts of near-surface outputs from these eight models over Belgium and (ii) to evaluate their suitability for driving a simple agricultural decision-making tool to help farmers identify optimal spraying windows for plant protection products.

2. Materials and Methods

2.1. The Network of Weather Stations

The study was conducted from 1 March 2025 to 31 August 2025. Weather observations served as the reference for evaluating forecast performance. The observations used in

this study were recorded by the automated weather stations of the PAMESEB network, managed by the Walloon Agricultural Research Centre (CRA-W). The PAMESEB network comprised 35 weather stations distributed across Wallonia (Belgium), a region occupying the southern half of the country. This study focused on 22 stations (Figure 1), located in rural areas and selected for their good siting, i.e., the absence of obstacles or heat sources that could influence the measurements, in accordance with the siting recommendations of the World Meteorological Organization. Each station consisted of an instrument enclosure and several sensors installed on or near a mast. The enclosure housed an OTT netDL 1000 data logger (OTT HydroMet GmbH, Kempten, Germany), which recorded data from all the sensors every minute. These data were transmitted via GPRS every hour to a central server. Air temperature and relative humidity were measured using a HMP155 probe (Vaisala, Vantaa, Finland), placed in a naturally ventilated meteoShield Professional radiation shield (Barani Design Technologies s.r.o., Bratislava, Slovakia). The radiation shield was mounted at a height of 1.5 m on the mast. The probe was recalibrated approximately once a year by the manufacturer or an accredited company. Global radiation was recorded using a SP Lite2 pyranometer (Kipp & Zonen B. V., Delft, The Netherlands) positioned at the top of the mast, at a height of 2 m. The pyranometer was recalibrated every two years by the manufacturer. Wind speed was measured by an A100R/K cup anemometer (Windspeed Ltd., Rhyl, UK) at a height of 2 m. At the beginning of the study period, most stations were equipped with an unheated YOUNG Model 52203 tipping-bucket rain gauge (R. M. YOUNG Company, Traverse City, MI, USA), with the top of the funnel positioned at a height of 1.2 m on a separate mast. This rain gauge was calibrated in our laboratory every six months. The only stations equipped with a weighing rain gauge before the start of the study period were Gembloux and Jevoumont. The weighing rain gauge was an OTT Pluvio (OTT HydroMet GmbH, Kempten, Germany). During the study period, nine additional stations were equipped with this rain gauge model: Ath (2 June), Casteau (20 May), Haut-le-Wastia (11 June), Libramont (17 June), Michamps (26 August), Ragnies (23 May), Seny (12 June), and Upigny (20 March). Data quality was ensured through semi-automatic quality control, combining automatic checks [17] with daily human supervision. This process helped to maintain a complete historical record of high-quality observations.

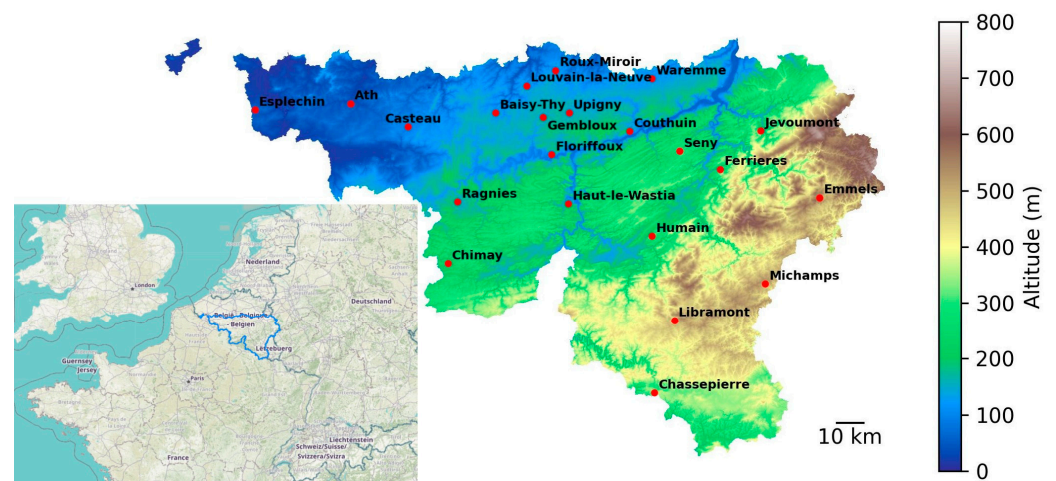


Figure 1. The 22 PAMESEB stations that were selected for the study. The location of Wallonia in Europe is indicated by a blue shape on the map in the bottom left corner. The locations of the stations in Wallonia are indicated by red dots on the topographic map. The map of Europe comes from OpenStreetMap. The topographic map has been provided by EuroGeographics 2022 and is available for free at <https://www.mapsforeurope.org> (accessed on 13 December 2024).

2.2. Overview of Weather Conditions During the Study Period

Wallonia has a temperate climate. Owing to its location in the mid-latitudes, the region is influenced by both polar and tropical air masses. Winters are mild and wet, whereas summers are relatively cool. Precipitation is evenly distributed throughout the year. During the 2000–2024 period, annual precipitation totals ranged from 700 to 1300 mm, while annual mean temperatures ranged from 8 to 11 °C. The Ardennes and the High Fens regions, located at higher elevations, are colder and receive more precipitation.

The study period began with a very dry and sunny spring. March was extremely dry. Across the entire station network, the highest precipitation total recorded in March was only 12 mm, whereas the average March precipitation over the 2000–2024 period ranged from 50 to 100 mm, depending on the station. Temperatures in March were well above average. April was warm and sunny, further intensifying the drought, particularly in western Wallonia. May was also sunny and dry, although precipitation was close to average in the easternmost part of Wallonia. Warm and sunny conditions persisted in June. It was only in July that most of Wallonia experienced above-average rainfall, while temperatures remained close to normal. Finally, August was dry and marked by a heatwave that occurred in the middle of the month.

2.3. Weather Forecast Data

Eight weather forecast models were evaluated (Table 1). The first four were global models with relatively coarse spatial resolutions. The Integrated Forecasting System of the ECMWF (Cycle 49r1) provided a deterministic physics-based model, referred to as HRES, and its ensemble version, referred to as ENS. This ENS model was treated as a deterministic model by considering the mean of all ensemble members. ECMWF also provided the AIFS Single V1 data-driven model, based on deep learning [13]. The last global model was the Global Forecast System (GFS), operated by NOAA.

Table 1. Characteristics of the forecast models. * Note: the native step (temporal resolution) of the HRES, ENS and GFS models depended on the forecast lead time. This table shows the step considered for this study, i.e the minimum step available for the entire forecast horizon.

Model	Scale	Type	Runs (UTC)	Step (h)	Forecast Horizon (Days)	Horizontal Resolution (km lat × km lon)
HRES	Global	Physics-based	00 & 12	6 *	10	17 × 7.5
ENS	Global	Physics-based	00 & 12	6 *	10	17 × 7.5
AIFS	Global	AI	00 & 12	6	10	27 × 27
GFS	Global	Physics-based	00 & 12	6 *	10	27 × 27
MAR	Regional	Physics-based	00 & 12	1	10	5 × 5
AROME	Regional	Physics-based	00 & 12	1	2	2.5 × 2.5
ICON-D2	Regional	Physics-based	00 & 12	1	2	2 × 2
HARMONIE	Regional	Physics-based	00 & 12	1	2.5	2 × 2

Four regional physics-based models were also tested. AROME was operated by the French national weather service (Météo-France) and driven at its limits by the global model ARPEGE [18]. ICON-D2 was operated by the German national weather service (DWD) and driven by the global model ICON [19]. HARMONIE was operated by the Dutch national weather service (KNMI) and driven by HRES [20,21]. The MAR model (version 3.14.0) [22] was run by the University of Liège (Belgium) and driven by GFS. Additionally, the ERA5 reanalysis data [23] were compared with observations. ERA5 was included to benchmark the performance of the forecasts at 0 h lead time with that of a reference gridded product. Owing to its assimilation of various observations, ERA5 was expected to

provide an improved estimation of the state of the atmosphere. It should be noted that the observations from the PAMESEB network, which are used as a reference in this study, were not incorporated into ERA5.

The forecast data of HRES, ENS and AIFS were retrieved via an API from the ECMWF Meteorological Archival and Retrieval System (MARS) (<https://apps.ecmwf.int/archive-catalogue>, accessed on 27 November 2025). GFS data were retrieved from the NOAA website (<https://nomads.ncep.noaa.gov>, accessed on 27 November 2025). The remaining models were accessed via an API from the open-meteo.com platform [24]. The 00 UTC and 12 UTC runs were used for this study. Although some models provided forecasts beyond 10 days, the analyses were limited to a 10-day forecast horizon. Regional models had a much shorter forecast horizon. AROME and ICON-D2 provided forecasts up to 2 days ahead, and HARMONIE up to 2.5 days ahead. ERA5 reanalysis data were retrieved from the Climate Data Store (CDS) and had a horizontal resolution of $0.25^\circ \times 0.25^\circ$, which is equivalent to $27 \text{ km} \times 27 \text{ km}$.

Five meteorological variables were analysed: air temperature, relative humidity, precipitation, wind speed, and global radiation (Table 2). However, not all variables were available for each model. Air temperature forecasts were available from all models. Global radiation forecasts were analysed for HRES, MAR and AIFS. Relative humidity and precipitation forecasts could not be retrieved from ENS. Precipitation and wind speed forecasts were also unavailable for GFS. The comparison of the models with the observations was carried out at 11 specific forecast lead times: 0, 6, 12, 18, 24, 48, 72, 96, 120, 168, and 240 h.

Table 2. Summary of weather variables tested for each model.

Model	Air Temperature (°C)	Air Humidity (%)	Precipitation (mm)	Wind Speed (m s ^{−1})	Global Radiation (W m ^{−2})
HRES	✓	✓	✓	✓	✓
ENS	✓			✓	
AIFS	✓	✓	✓	✓	✓
GFS	✓	✓			
MAR	✓	✓	✓	✓	✓
AROME	✓	✓	✓	✓	
ICON-D2	✓	✓	✓	✓	
HARMONIE	✓	✓	✓	✓	

This study had to deal with missing data due to difficulties in retrieving some forecast runs. For MAR, missing data resulted from server problems. For the models retrieved from open-meteo.com, missing data resulted from delayed model runs. Missing data for each model and forecast run are summarised in Table A1. For ICON-D2 and AROME, the results at 48 h lead time are not shown for the 12 UTC runs because data were not available at this lead time from 1 March to 18 March inclusive.

2.4. Matching Observations and Forecasts

Several preprocessing steps were performed to enable a meaningful comparison between observations and weather forecast models. The first preprocessing step involved converting all model outputs into the same units as the observations. HRES provided dew point temperature rather than relative humidity. Therefore, a conversion from dew point to relative humidity was applied, following the formulation of Goff [25] for saturation vapour pressure. All models provided wind speed at 10 m above ground level, except

MAR, which provided wind speed at 2 m. Wind speed was converted from 10 m to 2 m using a logarithmic wind profile (Equation (1)).

$$WS_{2m} = WS_{10m} \frac{\log(\frac{2}{Z_0})}{\log(\frac{10}{Z_0})} \quad (1)$$

WS_{2m} and WS_{10m} (m s^{-1}) are wind speeds at 2 and 10 m, respectively. Z_0 (m) is the rugosity length and was approximated to 0.05.

The second preprocessing step was to assign each weather station to the nearest model grid cell. The third preprocessing step involved the temporal matching of observations and forecasts. For MAR, the model outputs represented hourly averages. For example, the temperature value at 00:30 UTC was the average temperature between 00:00 UTC and 01:00 UTC. This value was compared with the observed average hourly temperature. For all the other models, temperature, relative humidity, and wind speed were instantaneous values. These instantaneous forecasts were compared with one-minute observations, rather than with hourly average observations. For precipitation and global radiation, the comparisons were based on the total precipitation and the average global radiation over 24 h periods.

The fourth preprocessing step was to remove numerical artefacts in the data. For example, MAR occasionally produced precipitation amounts as low as 0.0001 mm. For both observations and forecasts, precipitation values below 0.1 mm and global radiation values below 1 W m^{-2} were set to zero. Daily precipitation totals above 120 mm were excluded because AIFS and HRES data sometimes contained unrealistic daily precipitation, reaching up to several thousand millimetres. Since the smallest of these artefacts was 125 mm, a threshold of 120 mm was adopted. No extreme but plausible observation approached this threshold. The highest daily precipitation total recorded during the study period was 54.13 mm.

The fifth preprocessing step consisted of removing pairs of observations and forecasts for which the observations were considered unreliable. These implausible observations were flagged during the PAMESEB network quality control process [17], which relied on manual and automatic checks against physical and climatological thresholds, as well as temporal consistency with previous observations, spatial consistency with neighbouring stations, and internal consistency between sensors at the same station. For temperature, relative humidity, and wind speed, only the hours during which observations were implausible were removed. For precipitation and global radiation, if observations were flagged as implausible for two hours or more within the same day (00:00–00:00 UTC or 12:00–12:00 UTC), the entire day was excluded. This was necessary because subsequent analyses were based on daily accumulated values for these variables.

Overall, relatively few observations were excluded. No wind speed observations were excluded. For air temperature, only two observations were excluded. For relative humidity, about 100 values were removed. The largest number of excluded observations concerned precipitation, mainly due to clogged tipping-bucket rain gauges. Precipitation data from the Chimay station were excluded for more than one month, from the start of March to 15 April, while those from the Louvain-la-Neuve station were deleted from 31 July to the end of August. Very few global radiation observations were excluded. The largest number of excluded observations concerned the station of Humain, with 4 days deleted in March.

2.5. Decision-Support Tool for Identifying Optimal Spraying Windows in Agriculture

The quality of weather forecasts was assessed based on their impact on the output of a decision-support tool, compared to the output of the same tool fed by weather observations. The tool was a simple model to determine the best hours for spraying based on four weather

variables: air temperature, relative humidity, wind speed, and precipitation. These weather variables impact both product efficacy and losses in the environment. The algorithm of this tool is presented in Figure 2. A score was assigned to each weather variable based on its value. The four scores were then summed to produce one of five spraying classes: optimal, correct, limit, avoid, or forbidden. The forbidden class was assigned only when wind speed was greater than or equal to 20 km h^{-1} , which was a legal threshold in Belgium in 2025. The thresholds for temperature, relative humidity, and wind speed were established by local experts from the PROTECT'eau asbl. For rain, some forecast models only provided rainfall totals over 6 h periods. The criteria therefore had to be defined by combining expert knowledge with this constraint. Considering the rainfall totals over a 6 h period makes sense because the leaves can remain wet for several hours. For some plant protection products, it is best to wait for the leaves to dry before spraying.

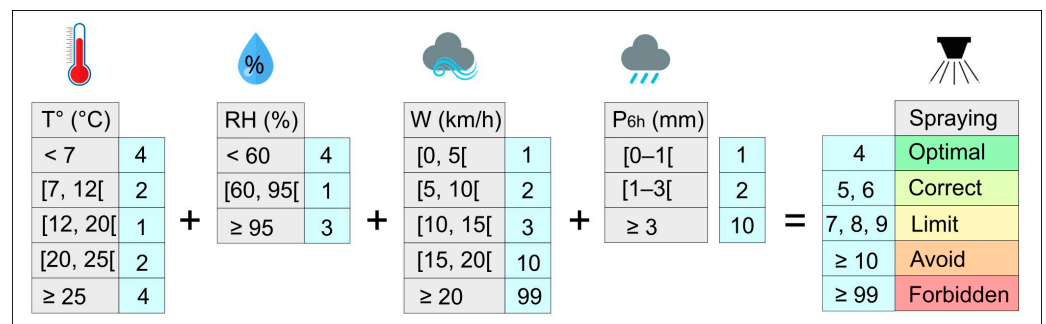


Figure 2. Criteria used to assess weather conditions for spraying. T° is the mean air temperature during the hour of interest ($^{\circ}\text{C}$), RH is the mean relative humidity during the hour of interest (%), W is the mean wind speed during the hour of interest (km h^{-1}), and P_{6h} is the cumulative precipitation over the 6 h period ending with the hour of interest (mm).

This threshold-based approach is consistent with many of the decision-support tools currently used by farmers. Nevertheless, the model is very simple, and it is important to acknowledge its main limitations. Firstly, the model only accounts for rainfall during a 6 h period ending at the hour of interest. It does not account for heavy rain in the next hours that could wash off the product. Nor does it consider rainfall during the preceding days that could have made the soil unsuitable for the passage of heavy machinery. Secondly, the model does not account for the fact that weather requirements vary according to the category of product being sprayed (e.g., root herbicide, foliar herbicide, plant growth regulator, fungicide) and the mode of action (e.g., contact or systemic). Thirdly, it does not account for temperature inversion that could lead to spray drift, even if the wind speed criterion is met. Finally, this is a threshold-based model. A small change in the input values can cause a threshold to be crossed and result in a very different output.

2.6. Performance Metrics

The comparison of weather forecast models was restricted to run dates for which all models had complete data. The number of excluded dates varied across analyses because the set of models being compared depended on the variables and forecast lead times considered. Of the 184 dates in the study period, up to 71 were excluded for the 00 UTC run and up to 46 for the 12 UTC run. Although this approach involved sacrificing a substantial amount of data, it ensured that all models were compared using the same set of observations.

Two metrics were used for the quantitative comparison between the observations and the forecasts of air temperature, relative humidity, wind speed, average global radiation over 24 h, and total precipitation over 24 h. The Mean Bias Error (MBE) indicates the

direction of model bias (underestimation or overestimation). The Root Mean Square Error (RMSE) provides an assessment of overall model performance. These indicators were stratified by month and by station. The same stratification was applied to the metrics presented in the next paragraphs. The RMSE was chosen instead of the Mean Absolute Error (MAE) because it gives more weight to large errors, which must be avoided to prevent poor agricultural decisions and costly consequences. For quantitative precipitation forecasts, the RMSE was calculated using only observation–forecast pairs for which the observed or forecast precipitation amount was greater than or equal to 0.1 mm.

$$MBE = \frac{1}{n} \sum_{i=1}^n (f_i - o_i) \quad (2)$$

With f the forecast and o the observation. The letter n represents the number of observations.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (f_i - o_i)^2}{n}} \quad (3)$$

A qualitative verification of precipitation forecasts was also performed. Precipitation was treated as a binary event and evaluated using a contingency table. A precipitation event was defined as daily precipitation greater than or equal to 1 mm. True positives (TPs), or hits, were events that were both observed and forecast. False positives (FPs), or false alarms, were events that were forecast but not observed. False negatives (FNs), or misses, were events that were observed but not forecast. True negatives (TNs) corresponded to occasions when no event was either observed or forecast. Although numerous metrics can be derived from a contingency table, this study focused on four metrics commonly used for the verification of precipitation forecasts. The description of these scores provided below is largely based on the documentation produced by the Joint Working Group on Forecast Verification Research, available on the knowledge hub of the World Meteorological Organization (accessed on 3 December 2025). The Equitable Threat Score (ETS), also called Gilbert Skill Score (GSS), calculates the proportion of events that were correctly predicted. It is adjusted for hits associated with random chance, to account for the fact that it is generally easier to forecast precipitation events in a wet climate than in a dry climate. It answers the question: how well did the forecast “yes” events correspond to the observed “yes” events, after accounting for hits due to chance? However, the ETS does not indicate whether there were more false alarms or misses. This information was provided by the Frequency Bias Index (FBI), the Probability of Detection (POD), and the False Alarm Ratio (FAR). The FBI compares the frequency of forecast events with the frequency of observed events. It expresses the bias of the model. The perfect value is 1. Values greater than 1 indicate that the model over-forecasts the event, whereas values lower than 1 indicate that the model under-forecasts the event. The POD measures the proportion of observed events that were forecast, whereas the FAR measures the fraction of forecast events that were in fact false alarms. These two metrics are easy to interpret for decision-makers.

$$ETS = \frac{TP - TP_{chance}}{TP + FP + FN - TP_{chance}}; TP_{chance} = \frac{(TP + FP)(TP + FN)}{TP + FP + FN + TN} \quad (4)$$

$$FBI = \frac{(TP + FP)}{(TP + FN)} \quad (5)$$

$$POD = \frac{TP}{TP + FN} \quad (6)$$

$$FAR = \frac{FP}{TP + FP} \quad (7)$$

The output of the spraying decision-support tool was one of the following five spraying condition classes: optimal, correct, limit, avoid, and forbidden. Therefore, the evaluation of the agreement between the outputs of the tool fed by the observations and the tool fed by the forecasts required a metric adapted for multiple ordinal classes. The metric selected in this study was the weighted kappa coefficient. This metric provided a quantitative measure of the level of agreement between spraying classes derived from observations and forecasts by subtracting the proportion of agreement expected by chance. The weighted kappa gives more weight to large classification errors [26]. For example, a confusion between “optimal” and “avoid” classes has more weight than a confusion between “optimal” and “correct” classes. Quadratic weighting was chosen to emphasise the penalisation of these large errors. The weighted kappa was computed using the Python sklearn.metrics library, version 1.4.2.

To determine whether two models differed significantly for a given metric, a bootstrap procedure was used to estimate the confidence interval of the difference in metric scores between the two models. For each lead time, bootstrapping was carried out for all possible model pairs. For each pair, 1000 bootstrap samples were generated by randomly sampling with replacement from the original dataset, and the same resampled observations were used to recompute the metric score for each model in the pair. Resampling was performed completely at random and was not stratified by station. The bootstrap distribution of score differences was obtained by calculating the difference in metric scores between the two models for each bootstrap sample. The 95% confidence interval of the difference between the scores of two models was obtained from the 2.5th and 97.5th percentiles of this bootstrap distribution. The scores of the two models were considered significantly different if this confidence interval did not include 0. Models were then assigned letter rankings according to the significance of their pairwise comparison. A model could receive multiple letters if it was not significantly different from the best-performing model within each letter group.

3. Results

3.1. Air Temperature

Table 3 and Table A2 present the RMSE of air temperature for the 12 UTC and 00 UTC runs, respectively. For lead times of less than 2 days, ICON-D2 and AIFS achieved the best performance. For lead times of 2 to 6 days, AIFS performed best, followed by ENS. For lead times of 7 days or more, ENS had a slight advantage over AIFS.

Table 3. RMSE (°C) of air temperature forecasts for the 12 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, orange, and red correspond to a, b, c, and d, respectively. Subsequent letters are also shown in red.

Model	Forecast Lead Time (h) and Corresponding UTC Hour											
	0 12	6 18	12 00	18 06	24 12	36 00	48 12	72 12	96 12	120 12	168 12	240 12
ERA5	1.23 ^c											
AROME	1.06 ^b	1.18 ^c	1.62 ^c	1.32 ^c	1.52 ^b	1.69 ^d						
ICON-D2	0.72 ^a	0.95 ^a	1.36 ^a	1.18 ^b	1.34 ^a	1.44 ^{ab}						
HARMONIE	1.22 ^c	1.36 ^d	1.66 ^c	1.34 ^c	1.64 ^c	1.72 ^d	1.88 ^d					
MAR	1.93 ^e	1.61 ^e	2.92 ^e	1.94 ^e	1.97 ^d	2.85 ^f	2.07 ^e	2.39 ^d	2.59 ^d	3.21 ^d	3.81 ^c	
HRES	1.04 ^b	1.41 ^d	1.44 ^b	1.13 ^{ab}	1.45 ^b	1.52 ^c	1.62 ^c	1.96 ^c	2.15 ^c	2.48 ^c	3.56 ^b	4.90 ^c
ENS	1.03 ^b	1.4 ^d	1.38 ^a	1.11 ^a	1.40 ^{ab}	1.46 ^b	1.55 ^b	1.76 ^b	1.98 ^b	2.21 ^b	3.14 ^a	4.04 ^a
GFS	1.67 ^d	1.88 ^f	1.89 ^d	1.47 ^d	2.00 ^d	2.00 ^e	2.21 ^f	2.68 ^e	2.94 ^e	3.45 ^e	3.93 ^c	5.71 ^d
AIFS	1.02 ^b	1.04 ^b	1.37 ^a	1.08 ^a	1.32 ^a	1.40 ^a	1.39 ^a	1.62 ^a	1.82 ^a	2.10 ^a	3.18 ^a	4.5 ^b

The MAR model made large errors at 00:00. For the first few days of lead time, it strongly overestimated air temperature at 00:00 by slightly less than 2 °C on average

(Figure A1). Furthermore, MAR underestimated the air temperature at 12:00 by slightly less than 1 °C. GFS was also biased. It underestimated air temperature at 00:00 by slightly less than 1 °C on average and overestimated it at 12:00 by approximately 0.5 °C. The biases did not increase much with the lead time. Overall, the RMSE did not differ much between weather stations. However, at 00:00, all models had a significantly higher RMSE at the Ferrières station and the Floriffoux station, located in a village in a river valley. AROME was one of the models that overestimated the air temperature the most at 00:00 in Floriffoux. The overestimation was +2 °C on average but reached as high as +8 °C.

MAR showed a positive bias of approximately 3 °C at 00:00 from May to August (Figure A2). GFS had a negative bias throughout the study period, and it was most pronounced around May. At 12:00, MAR bias was very negative in spring (about −1.5 °C), and GFS bias was very positive in July and August (about +1.5 °C).

3.2. Relative Humidity

Table 4 and Table A3 present the RMSE of relative humidity forecasts for the 12 UTC and 00 UTC runs, respectively. AIFS was the best-performing model across the entire forecast horizon, except at 0 h lead time. ICON-D2 also performed well for lead times less than 2 days. MAR strongly overestimated relative humidity, particularly at 12:00. This overestimation was greatest in March, with an average overestimation slightly above 20% at 12:00. For the 12 UTC runs, AIFS and ICON-D2 had a lower RMSE than ERA5, but this was not the case for the 00 UTC runs.

Table 4. RMSE (%) of relative humidity forecasts for the 12 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, orange, and red correspond to a, b, c, and d, respectively. Subsequent letters are also shown in red.

Model	Forecast Lead Time (h) and Corresponding UTC Hour											
	0 12	6 18	12 00	18 06	24 12	36 00	48 12	72 12	96 12	120 12	168 12	240 12
ERA5	6.7 ^c											
AROME	7.3 ^d	9.7 ^c	11.1 ^d	8.7 ^c	9.9 ^c	12.5 ^e						
ICON-D2	5.4 ^a	7.3 ^b	8.4 ^b	7.0 ^{ab}	8.9 ^b	9.3 ^b						
HARMONIE	8.6 ^e	12.4 ^e	10.0 ^c	8.5 ^c	9.8 ^c	10.9 ^d	11.0 ^b					
MAR	18.3 ^g	17.1 ^g	13.4 ^e	10.1 ^d	18.0 ^e	13.3 ^f	18.3 ^d	19.1 ^d	19.7 ^d	21.5 ^d	22.4 ^d	
HRES	7.0 ^d	10.8 ^d	8.8 ^b	7.2 ^b	9.2 ^b	9.9 ^c	10.9 ^b	11.8 ^b	12.9 ^b	14.7 ^b	16.8 ^b	19.5 ^b
GFS	11.2 ^f	14.9 ^f	11.5 ^d	11.6 ^e	12.8 ^d	12.4 ^e	13.9 ^c	15.7 ^c	18.0 ^c	19.4 ^c	20.9 ^c	23.5 ^c
AIFS	5.9 ^b	6.8 ^a	7.8 ^a	6.8 ^a	8.4 ^a	8.8 ^a	9.1 ^a	9.8 ^a	10.7 ^a	12.0 ^a	15.8 ^a	17.7 ^a

For all models, the RMSE was higher at 00:00 at the Ferrières and Floriffoux stations. At 00:00 in Floriffoux, AROME had the highest RMSE and the strongest negative bias (approximately 14%) (Figure A3). However, at 12:00, the RMSE of the models was constant across stations. The models had a slightly positive bias at 00:00 in spring, but this bias became slightly negative in summer (Figure A4). GFS had a more pronounced negative bias at 12:00 in July and August.

3.3. Wind Speed

Table 5 and Table A4 present the RMSE of wind speed for the 12 UTC and 00 UTC runs, respectively. ICON-D2 had the best performance for the first two days of lead time. After 4–5 days of lead time, ENS clearly outperformed the other models.

Interestingly, the errors did not always increase with lead time. For the 12 UTC runs, the RMSE at 6, 12, 18, and 36 h lead times (corresponding to 18:00, 00:00 and 06:00) was lower than the RMSE at 0 and 24 h lead times (corresponding to 12:00), for all models

except MAR. This is correlated with lower wind speed at night and higher wind speed during the day. The average wind speed over the entire study period and across all stations at night (from 18:01 to 06:00) was 1.50 m s^{-1} , and the average during the day (from 6:01 to 18:00) was 2.62 m s^{-1} . Unlike other models, the RMSE of MAR did not change much for the lead times from 6 to 72 h. This was observed for both the 12 UTC and 00 UTC runs. MAR forecast wind speed well at 12:00, but for 06:00, 18:00 and 00:00 it was one of the worst-performing models.

Table 5. RMSE (m s^{-1}) of wind speed forecasts for the 12 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, orange, and red correspond to a, b, c, and d, respectively. Subsequent letters are also shown in red.

Model	Forecast Lead Time (h) and Corresponding UTC Hour											
	0 12	6 18	12 00	18 06	24 12	36 00	48 12	72 12	96 12	120 12	168 12	240 12
ERA5	0.99 ^a											
AROME	1.25 ^c	1.03 ^{ab}	0.93 ^{bc}	0.90 ^{ab}	1.28 ^a	0.99 ^c						
ICON-D2	1.2 ^b	0.99 ^a	0.81 ^a	0.87 ^a	1.26 ^a	0.88 ^a						
HARMONIE	1.46 ^e	1.05 ^b	0.91 ^b	0.89 ^{ab}	1.43 ^b	0.94 ^b	1.47 ^b					
MAR	1.31 ^d	1.36 ^c	1.35 ^e	1.23 ^d	1.28 ^a	1.38 ^d	1.27 ^a	1.30 ^a	1.42 ^b	1.49 ^b	1.70 ^b	
HRES	1.28 ^{cd}	1.04 ^b	0.95 ^c	0.92 ^b	1.29 ^a	0.96 ^{bc}	1.32 ^a	1.40 ^b	1.45 ^b	1.51 ^b	1.71 ^b	1.83 ^b
ENS	1.28 ^{cd}	1.00 ^{ab}	0.94 ^{bc}	0.91 ^b	1.25 ^a	0.94 ^b	1.27 ^a	1.32 ^a	1.35 ^a	1.41 ^a	1.45 ^a	1.49 ^a
AIFS	1.3 ^{cd}	1.00 ^{ab}	0.99 ^d	0.96 ^c	1.27 ^a	0.98 ^c	1.32 ^a	1.38 ^b	1.43 ^b	1.51 ^b	1.68 ^b	1.85 ^b

For the 12 UTC runs, ERA5 had the lowest RMSE, whereas for the 00 UTC runs, ERA5 had the second-highest RMSE.

At 00:00, the models had a negative bias of around -1.5 m s^{-1} at the Ragnies station and a positive bias of around 0.5 m s^{-1} at the Esplechin station (Figure A5). It should be noted that the Ragnies station was the windiest station during the day, with an observed average wind speed of 3.83 m s^{-1} . There was no significant change in RMSE over the months for any of the models.

3.4. Precipitation

Table 6 and Table A5 present the RMSE of daily precipitation for the 12 UTC and 00 UTC runs, respectively. AIFS had the numerically lowest RMSE value at every lead time, although it was not significantly different from at least one other model at most lead times.

Table 6. RMSE (mm) of daily precipitation forecasts ($\geq 0.1 \text{ mm}$) for the 12 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, and orange correspond to a, b, and c, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour						
	24 12	48 12	72 12	96 12	120 12	168 12	240 12
AROME	5.0 ^{bc}						
ICON-D2	4.8 ^b						
HARMONIE	5.6 ^c	6.2 ^c					
MAR	4.3 ^{ab}	5.1 ^b	4.6 ^{ab}	5.5 ^b	5.5 ^a	5.5 ^a	
HRES	4.5 ^{ab}	4.4 ^a	5.0 ^b	5.1 ^b	6.6 ^b	6.8 ^b	7.1 ^b
AIFS	3.9 ^a	4.0 ^a	4.0 ^a	4.2 ^a	5.1 ^a	5.1 ^a	6.2 ^a

Figure 3 shows the monthly RMSE of daily precipitation for the 12 UTC runs. The RMSE was lowest in March, increased in spring, and peaked in July. March was a very dry

month, which may explain the low RMSE. The maximum reached in July likely corresponded to a peak in thunderstorm activity. The station of Seny stood out from the other stations with a higher RMSE at every lead time. This could be explained by the underestimation of precipitation by all models for the series of thunderstorms that struck this station in July.

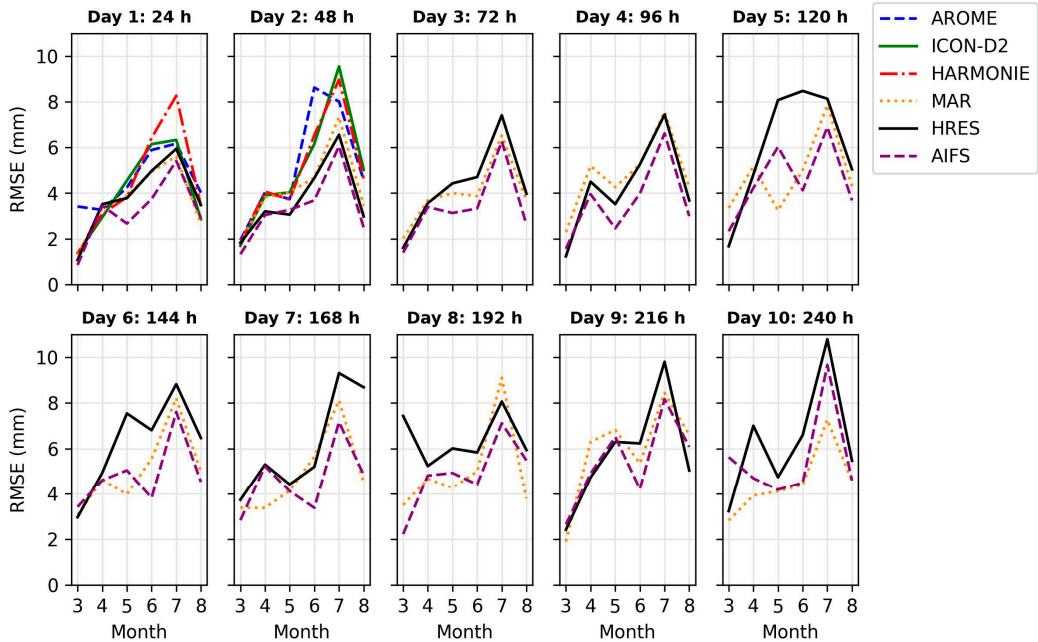


Figure 3. Monthly RMSE of daily precipitation (≥ 0.1 mm) for each forecast lead time of the 12 UTC runs.

The ETS metric was used to evaluate the ability of the models to predict binary precipitation occurrence: rainy day or non-rainy day. The same conclusions were drawn for the 12 UTC and the 00 UTC runs, presented in Table 7 and Table A6, respectively. AROME and ICON-D2 were among the best models for 24 h lead time. For lead times beyond 48 h, HRES and AIFS had the best ETS values. MAR was the poorest-performing model for predicting daily precipitation occurrence.

Table 7. ETS of rainy day (daily precipitation ≥ 1 mm) event forecasts for the 12 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, and orange correspond to a, b, and c, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour						
	24	48	72	96	120	168	240
	12	12	12	12	12	12	12
AROME	0.53 ^{ab}						
ICON-D2	0.52 ^{ab}						
HARMONIE	0.50 ^b	0.49 ^a					
MAR	0.39 ^c	0.36 ^c	0.37 ^b	0.29 ^b	0.19 ^b	0.12 ^b	
HRES	0.56 ^a	0.44 ^b	0.43 ^a	0.38 ^a	0.35 ^a	0.18 ^a	0.11 ^b
AIFS	0.54 ^{ab}	0.45 ^{ab}	0.41 ^a	0.34 ^a	0.32 ^a	0.21 ^a	0.16 ^a

Table 8 and Table A7 present the False Alarm Ratio (FAR) for the 12 UTC and 00 UTC runs, respectively. The FAR increased with lead time, except for the first three lead times for MAR. The FAR was higher in March for all models, and at every lead time, it reached values between 50% and 80% at the first three lead times for every model.

Table 8. FAR (%) of rainy day (daily precipitation ≥ 1 mm) event forecasts (rain or no-rain) for the 12 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, and orange correspond to a, b, and c, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour						
	24 12	48 12	72 12	96 12	120 12	168 12	240 12
AROME	32 ^b						
ICON-D2	30 ^{ab}						
HARMONIE	35 ^b	33 ^a					
MAR	46 ^c	48 ^c	46 ^a	52 ^b	62 ^c	68 ^b	
HRES	26 ^a	40 ^b	42 ^a	46 ^a	47 ^a	62 ^a	64 ^a
AIFS	32 ^b	43 ^b	46 ^a	52 ^b	53 ^b	62 ^a	62 ^a

Table 9 and Table A8 present the Probability of Detection (POD) for the 12 UTC and 00 UTC runs, respectively. AIFS had the best POD for every lead time. MAR had the lowest POD for lead times beyond 48 h. POD values were lowest in March, with the best-performing model (AIFS) achieving a POD of only 70%.

Table 9. POD (%) of rainy day (daily precipitation ≥ 1 mm) event forecasts (rain or no-rain) for the 12 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, and orange correspond to a, b, and c, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour						
	24 12	48 12	72 12	96 12	120 12	168 12	240 12
AROME	87 ^a						
ICON-D2	82 ^b						
HARMONIE	86 ^a	82 ^c					
MAR	89 ^a	85 ^{bc}	82 ^c	73 ^c	67 ^c	52 ^c	
HRES	82 ^b	86 ^b	86 ^b	82 ^b	76 ^b	61 ^b	53 ^b
AIFS	88 ^a	96 ^a	94 ^a	94 ^a	89 ^a	74 ^a	73 ^a

Table 10 and Table A9 present the Frequency Bias Index (FBI) for the 12 UTC and 00 UTC runs, respectively. For both runs, all models overforecast precipitation occurrence (FBI > 1) at every lead time, and the FBI increased with lead time. For all models, FBI values were substantially higher in March than in the other months.

Table 10. FBI of rainy day (daily precipitation ≥ 1 mm) event forecasts (rain or no-rain) for the 12 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, and orange correspond to a, b, and c, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour						
	24 12	48 12	72 12	96 12	120 12	168 12	240 12
AROME	1.28 ^b						
ICON-D2	1.16 ^a						
HARMONIE	1.32 ^b	1.21 ^a					
MAR	1.64 ^c	1.64 ^c	1.52 ^a	1.51 ^a	1.74 ^b	1.63 ^a	
HRES	1.10 ^a	1.44 ^b	1.49 ^a	1.52 ^a	1.43 ^a	1.60 ^a	1.47 ^a
AIFS	1.3 ^b	1.68 ^c	1.73 ^b	1.96 ^b	1.9 ^c	1.93 ^b	1.94 ^b

3.5. Global Radiation

Table 11 and Table A10 present the RMSE of mean daily global radiation for the 12 UTC and 00 UTC runs, respectively. AIFS achieved the lowest RMSE at every lead time.

Table 11. RMSE (W m^{-2}) of mean daily global radiation for the 12 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, and orange correspond to a, b, and c, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour						
	24 12	48 12	72 12	96 12	120 12	168 12	240 12
ERA5	28.5 ^b						
MAR	38.4 ^c	40.4 ^c	43.7 ^c	47.5 ^c	56.3 ^c	67.9 ^c	
HRES	16.5 ^a	32.5 ^b	35.7 ^b	41.7 ^b	46.5 ^b	61.7 ^b	93.4 ^b
AIFS	16.2 ^a	28.3 ^a	32.5 ^a	35.1 ^a	42.5 ^a	54.2 ^a	86.5 ^a

HRES overestimated global radiation at all stations and at every lead time, while AIFS, MAR and ERA5 underestimated it. The bias remained relatively constant across lead times.

3.6. Case Study I: Incoming Thunderstorms

This section details a case study in which the RMSE of air temperature was high for most of the models. It happened on 14 June. Figure 4 shows the observed air temperature at each station at 12:00 together with the corresponding forecast air temperature from each model, based on the 12 UTC run of 13 June. AROME overestimated air temperature at all stations, except in Emmels, with the largest errors occurring in central and southern Wallonia. ICON-D2 overestimated air temperature at all the stations, with strong overestimation in central Wallonia. AIFS strongly overestimated air temperature in almost every weather station.

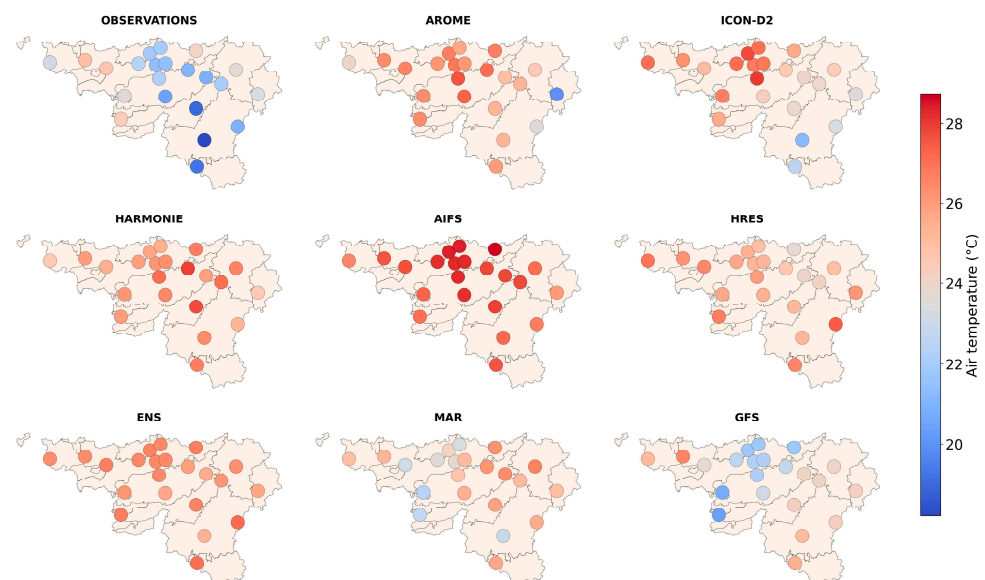


Figure 4. Maps of air temperature for 14 June at 12:00. The upper left map presents the observed temperatures, and the other maps present the temperatures forecast by 8 different models. Each bubble indicates the location of a weather station. The colour scale ranges from the lowest to the highest temperature, among all the observed or forecast temperatures for that hour. The darker the red, the higher the temperature. The darker the blue, the lower the temperature.

Figure 5 shows the observed 24 h precipitation totals between 13 June and 14 June at 12:00, as well as the forecast precipitation totals. No rainfall was observed on 13 June, but MAR forecast moderate continuous rainfall at every station for several hours during the afternoon. The other models performed quite well for this afternoon. During the night, a band of thunderstorms approached the west of the country but deteriorated rapidly. The Esplechin station, located at the western edge of the study area, recorded 7 mm and the next station to the east recorded 2 mm. MAR failed to detect this rainfall while the other models captured the event. However, most of them overestimated the rainfall amounts, and several predicted the precipitation band too far east.

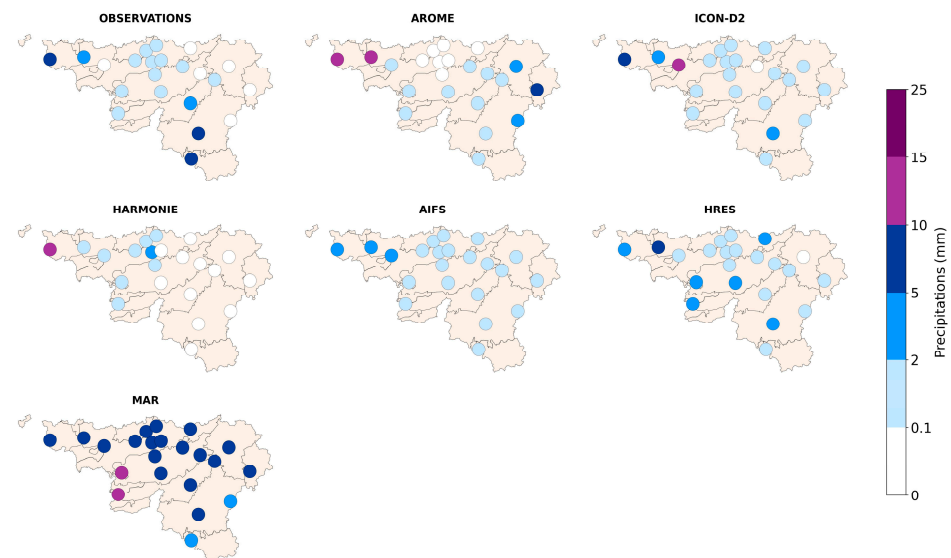


Figure 5. Maps of precipitation from 13 June at 12:00 to 14 June at 12:00. The upper left map presents the observed precipitation; the other maps present the precipitation forecast by 8 different models. Each bubble indicates the location of a weather station.

During the morning of 14 June, very light rainfall (<1 mm) was recorded at some weather stations in central Wallonia. Around 11:00, thunderstorms developed near the French border and struck the southernmost part of the region, including the Libramont and Chassepierre stations. All the models failed to predict this event correctly. AROME forecast thunderstorms further east but failed to predict any over southern Wallonia. ICON-D2 forecast a thunderstorm over Casteau during the morning, although no precipitation was observed there. This forecast was too far east of the observed precipitation area, which at that time was limited to the far west of the region. ICON-D2 also underestimated thunderstorm activity over Libramont and Chassepierre at noon. HARMONIE did not predict any thunderstorm activity on 14 June. MAR forecast rainfall over much of southern and eastern Wallonia during the afternoon. Because this rainfall occurred after 12:00 UTC, it was not included in the 24 h accumulated totals. Timing errors were not taken into account. HRES accurately predicted the rainy area during the night of 13–14 June. It also detected the thunderstorm that developed around noon over Libramont, although it did not detect it in Chassepierre. Finally, AIFS forecast light rain throughout the night and the following morning (approximately 1–2 mm maximum).

3.7. Case Study II: Heat Spike in July

This case study allows for an examination of model performance in forecasting a period of unusually high temperatures. Indeed, the first two days of July were particularly hot, with maximum temperatures exceeding 35 °C in some areas. On 3 July, temperatures dropped considerably (maximums around 25 °C). The RMSE values of the models were computed for the 0, 12, 24, 36, and 48 h lead times, for the 12 UTC runs of 30 June and

1 July. On average, across the lead times of the 12 UTC run of 30 June, the models with the lowest RMSE were ICON-D2, AIFS, HRES and ENS. The models with the highest RMSE were MAR, AROME, and HARMONIE. MAR had a particularly high RMSE, especially at 00:00. HARMONIE's temperature forecast at the 48 h lead time was significantly lower than the observed temperature, with differences of more than 10 °C for some stations. For the 12 UTC run of 1 July, the RMSE for all the models, except for GFS, showed that the temperature forecasts at 48 h lead time (3 July) were much more reliable than the forecasts at 12 h lead time (2 July, during the heat spike). Figure 6 shows the dynamics of observed and forecast air temperature for the weather station of Couthuin.

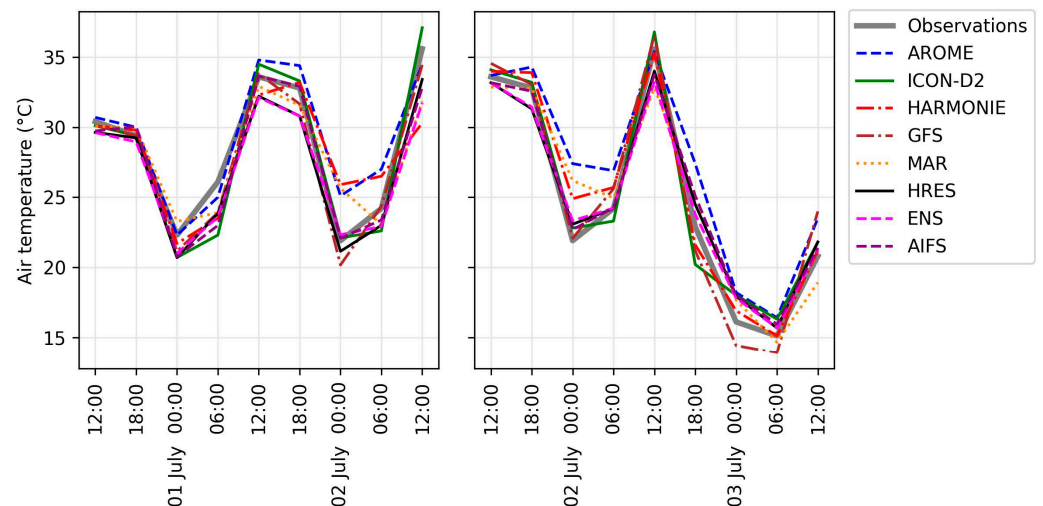


Figure 6. Dynamics of observed and forecast air temperature in Couthuin for the 12 UTC runs of 30 June and 1 July.

3.8. Case Study III: Heat Wave in August

This case study is another examination of model performance in forecasting a period of high temperatures. From 10 to 15 August, a large part of the network's weather stations experienced a heatwave, defined by five consecutive days with a maximum temperature reaching 25 °C, including three days reaching 30 °C. The RMSE of the forecast models was calculated at several lead times for the 12 UTC runs of 12 August (Table 12) and 13 August (Table 13). AIFS was the best-performing model for predicting air temperature during this heatwave event. MAR and GFS had a strong positive temperature bias.

Table 12. RMSE of air temperature (°C) for the 12 UTC run of 12 August. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, orange, and red correspond to a, b, c, and d, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour				
	0 12	12 00	24 12	36 00	48 12
ERA5	1.95 ^c				
AROME	1.16 ^b	2.00 ^{ab}	0.93 ^b	2.50 ^{bc}	2.54 ^{bc}
ICON-D2	0.63 ^a	1.52 ^a	0.51 ^a	2.42 ^{bc}	2.29 ^b
HARMONIE	1.01 ^b	1.71 ^{ab}	1.03 ^{bc}	1.24 ^a	2.31 ^b
MAR	1.81 ^c	4.22 ^c	1.40 ^c	3.54 ^c	3.45 ^c
HRES	1.12 ^b	1.50 ^a	0.98 ^{bc}	1.85 ^b	2.08 ^{ab}
ENS	1.09 ^b	1.37 ^a	1.05 ^{bc}	1.41 ^{ab}	1.70 ^{ab}
GFS	3.08 ^d	2.16 ^b	3.17 ^d	3.00 ^c	4.23 ^c
AIFS	1.03 ^b	1.53 ^a	0.52 ^a	1.11 ^a	1.58 ^a

Table 13. RMSE of air temperature (°C) for the 12 UTC run of 13 August. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, orange, and red correspond to a, b, c, and d, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour				
	0 12	12 00	24 12	36 00	48 12
ERA5	1.43 ^c				
AROME	0.85 ^b	2.53 ^c	1.91 ^{ab}	1.12 ^a	1.13 ^a
ICON-D2	0.50 ^a	2.09 ^c	1.57 ^a	1.98 ^{bc}	1.98 ^{bc}
HARMONIE	0.96 ^b	2.32 ^c	1.96 ^{ab}	2.53 ^c	1.34 ^{ab}
MAR	1.09 ^{bc}	3.78 ^d	2.37 ^b	4.16 ^d	2.31 ^c
HRES	1.10 ^{bc}	1.53 ^b	3.10 ^{bc}	1.76 ^b	1.88 ^{bc}
ENS	1.11 ^{bc}	1.37 ^{ab}	2.70 ^b	1.51 ^{ab}	1.63 ^{ab}
GFS	2.08 ^d	3.40 ^d	3.76 ^c	1.27 ^{ab}	4.38 ^d
AIFS	0.70 ^{ab}	1.08 ^a	1.52 ^a	1.02 ^a	1.83 ^b

3.9. Evaluation of Weather Forecasts for Identifying Optimal Spray Time Windows

The weighted kappa score measured the consistency between the output of the decision-support tool fed by the observations and the output of the tool fed by the forecast models. This kappa score was computed for each model and for each forecast lead time. For both the 12 UTC runs (Table 14) and the 00 UTC runs (Table 15), ICON-D2 and AIFS ranked first up to 48 h lead time. AROME ranked second. For lead times beyond 48 h, AIFS was the best-performing model. For all models, the weighted kappa was generally higher at 00:00 and 06:00 than at 12:00 on the same day. It was true even at long lead times. The MAR model, however, was proportionally less accurate than the others at 00:00 for the 12 UTC runs. It may be related to its poor performance in estimating wind speed at night for these runs.

Table 14. Weighted kappa indicating the ability of weather forecasts to correctly predict spraying classes. A kappa of 1 corresponds to a perfect prediction. Forecasts were computed during the 12 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, and orange correspond to a, b, and c, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour											
	6 18	12 00	18 06	24 12	30 18	36 00	42 06	48 12	54 18	60 00	66 06	72 12
AROME	0.46 ^b	0.56 ^b	0.57 ^b	0.35 ^b	0.43 ^b	0.51 ^b	0.53 ^b					
ICON-D2	0.53 ^a	0.65 ^a	0.61 ^{ab}	0.41 ^a	0.42 ^b	0.59 ^a	0.57 ^{ab}					
HARMONIE	0.37 ^c	0.58 ^b	0.59 ^{ab}	0.35 ^b	0.37 ^c	0.54 ^b	0.55 ^{ab}	0.32 ^b				
MAR	0.34 ^c	0.36 ^c	0.46 ^c	0.37 ^{ab}	0.35 ^c	0.33 ^c	0.44 ^c	0.37 ^{ab}	0.36 ^b	0.35 ^b	0.46 ^b	0.34 ^a
HRES	0.42 ^b	0.59 ^b	0.6 ^{ab}	0.41 ^a	0.36 ^c	0.55 ^{ab}	0.54 ^{ab}	0.34 ^b	0.34 ^b	0.57 ^a	0.55 ^a	0.3 ^a
AIFS	0.53 ^a	0.59 ^b	0.62 ^a	0.4 ^{ab}	0.49 ^a	0.54 ^{ab}	0.58 ^a	0.39 ^a	0.49 ^a	0.56 ^a	0.56 ^a	0.35 ^a
	102 18	108 00	114 06	120 12	150 18	156 00	162 06	168 12	222 18	228 00	234 06	240 12
MAR	0.24 ^b	0.21 ^b	0.39 ^b	0.28 ^a	0.18 ^b	0.13 ^c	0.26 ^b	0.13 ^b	0.05 ^b	0.1 ^c		
HRES	0.24 ^b	0.45 ^a	0.51 ^a	0.28 ^a	0.17 ^b	0.37 ^b	0.41 ^a	0.15 ^{ab}	0.17 ^a	0.31 ^b	0.29 ^a	0.01 ^b
AIFS	0.31 ^a	0.49 ^a	0.51 ^a	0.3 ^a	0.26 ^a	0.46 ^a	0.40 ^a	0.19 ^a	0.17 ^a	0.41 ^a	0.3 ^a	0.15 ^a

These weighted kappa values below 0.4 highlight poor consistency between forecasts and observations. For the 12 UTC runs and for all models, the score of the best model (AIFS) dropped below 0.4 at 48 h lead time for daytime forecasts (12:00), and at 102 h lead time for evening forecasts (18:00). For the 00 UTC runs, the score of the best model dropped below

0.4 at 60 h lead time for daytime forecasts, and at 114 h lead time for evening forecasts. For both runs, the scores for night-time (00:00) and morning (06:00) forecasts remained above 0.4 much longer. Overall, forecasts remained better than random chance for all the lead times, except for the 12 UTC runs at 240 h lead time where the score of HRES dropped close to 0.

Table 15. Weighted kappa indicating the ability of weather forecasts to correctly predict spraying classes. A kappa of 1 corresponds to a perfect prediction. Forecasts were computed during the 00 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, and orange correspond to a, b, and c, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour											
	6 06	12 12	18 18	24 00	30 06	36 12	42 18	48 00	54 06	60 12	66 18	72 00
AROME	0.64 ^a	0.44 ^a	0.44 ^b	0.50 ^b	0.56 ^b	0.43 ^a	0.39 ^b	0.37 ^b				
ICON-D2	0.68 ^a	0.48 ^a	0.46 ^b	0.59 ^a	0.57 ^{ab}	0.45 ^a	0.40 ^b	0.47 ^a				
MAR	0.46 ^b	0.43 ^a	0.37 ^c	0.33 ^c	0.44 ^c	0.39 ^a	0.36 ^{bc}	0.23 ^c	0.42 ^b	0.38 ^a	0.32 ^b	0.33 ^b
HRES	0.64 ^a	0.42 ^a	0.39 ^c	0.56 ^a	0.56 ^b	0.41 ^a	0.34 ^c	0.48 ^a	0.57 ^a	0.35 ^a	0.35 ^b	0.55 ^a
AIFS	0.66 ^a	0.43 ^a	0.51 ^a	0.50 ^b	0.61 ^a	0.42 ^a	0.48 ^a	0.44 ^a	0.57 ^a	0.37 ^a	0.43 ^a	0.57 ^a
	102 06	108 12	114 18	120 00	150 06	156 12	162 18	168 00	222 06	228 12	234 18	240 00
MAR	0.39 ^b	0.28 ^{ab}	0.21 ^b	0.23 ^c	0.32 ^c	0.21 ^b	0.20 ^b	0.15 ^c	0.26 ^b	0.17 ^a	0.15 ^a	0.16 ^c
HRES	0.54 ^a	0.27 ^b	0.22 ^b	0.46 ^b	0.39 ^b	0.19 ^b	0.17 ^b	0.36 ^b	0.31 ^a	0.11 ^b	0.19 ^a	0.37 ^a
AIFS	0.55 ^a	0.32 ^a	0.28 ^a	0.53 ^a	0.45 ^a	0.28 ^a	0.30 ^a	0.45 ^a	0.33 ^a	0.11 ^b	0.17 ^a	0.27 ^b

4. Discussion

4.1. Challenges Encountered in Conducting This Study

The first challenge was identifying open data sources and collecting forecasts for all the models. Forecast archives were rarely centralised, often heterogeneous, and sometimes incomplete. The second challenge was to automate the downloads of the forecasts to avoid missing some runs. This was not straightforward because the times of run publication could vary from day to day, sometimes with delays of several hours. These fluctuations led to erroneous downloads that had to be identified and removed from the dataset. For some models, such as HARMONIE, which updated its runs hourly, these delays resulted in the loss of many runs. The third challenge was data storage. Each run included a complete set of meteorological variables over the entire forecast horizon and for all grid cells. For example, a single run of the full grid of MAR, including all variables (not only those used in this study), produced approximately 150 MB of data. When multiplied by the two daily runs and the number of days in the study period, this reached 55 GB for this model alone. In comparison, HRES and ENS outputs required less storage space because the data were stored in a compact binary format. For HRES, all information covering the entire study period was contained in roughly 130 MB, as the selected grid was smaller and only the relevant variables were filtered. Finally, the fourth challenge was to implement preprocessing steps to make the data comparable to each other and to observations. This work included converting heterogeneous units, transforming and harmonising variables, but also rigorously mapping the spatiotemporal relationships between model outputs and observations.

4.2. Comparison with Other Studies

For wind speed, our study shows that the models made larger errors during the day, when wind speeds were higher. Häuser [6] also reported that models had more difficulty predicting wind speed when speeds were higher. The author noted that wind forecasting

was better in continental areas than near the sea and that ICON performed best among all the models tested, with similar or even better performance than ERA5 reanalysis data. In our study, ICON global was not tested. ICON-D2 had a better performance than ERA5 for the 00 UTC runs but not for the 12 UTC runs.

For global radiation, Reiß [7] obtained better estimates with ERA5 than with GFS or ICON. However, the author showed that combining several forecast models improves global radiation estimation. Such multi-models performed slightly better than ERA5. In our study, HRES and AIFS global radiation forecasts largely outperformed ERA5.

For precipitation, Haiden et al. [3] studied the performance of global models over an 11-month study period, using observations from all around the world. The ETS (threshold ≥ 1 mm) of HRES at 24 h lead time was just below 0.35. At 72 h lead time, ETS was between 0.3 and 0.25, and at 120 h it was about 0.2. The ETS values for the same model were much higher in our study, which may be due to the focus on Belgium and the improvement of HRES over time. The same authors also reported that the FBI (threshold ≥ 1 mm) of all the models was above 1 and did not change much with the lead time. In our study, the FBI of all the models was greater than 1 but tended to increase with lead time.

The results obtained from ECMWF products in this study were compared with 2024 Green Books from countries surrounding Belgium. The German Green Book noted the strong performance of AIFS compared to ICON and HRES. For synoptic stations in the Northern Hemisphere in April 2024, AIFS was the best-performing model for predicting air temperature, and this was true for most lead times from 24 h until 168 h. In our study, the magnitudes of the errors of AIFS and HRES were similar for air temperature at these lead times. For April 2024, AIFS seemed to have difficulty forecasting wind, especially at lead times up to 120 h. The errors were larger at 00:00 for AIFS, with much higher errors at 00:00 than at 12:00, and for HRES the error was stronger at noon. In our study, the errors were much lower for wind speed. Regarding 6 h precipitation totals, the German Green Book stated that AIFS provided good-quality forecasts, but with a strong bias at low precipitation rates. A bias (FBI) was also observed in our study.

The French Green Book provided an estimation of the RMSE of the ECMWF ensemble model (ENS) post-treated for maximum air temperature at 24, 48, 72 and 96 h lead times. The observations came from a network of weather stations in France. The report showed that the RMSE of maximum air temperature in July and August was about twice as high as in the other months. In our study, by contrast, almost no difference was found between the months for air temperature. In the French Green Book, considering a full year, the RMSE at 24 h lead time for maximum air temperature was between 1.5 and 2.0 °C. The RMSE increased with lead time and was between 2.0 and 2.5 °C at 96 h lead time. In our study, for air temperature, the RMSE of ENS was slightly lower.

The 2021 Belgian Green Book computed the RMSE of HRES and ENS over one year for the 12 UTC and 00 UTC runs. Forecast outputs were compared with observations from the synoptic station of Zaventem in Belgium. The RMSE of air temperature at 24 h lead time for each model was similar to the RMSE in our study. At 120 and 168 h lead times, ENS outperformed HRES, as in our study. At these lead times, the RMSE of each model was similar to the RMSE in our study for the 00 UTC runs, but lower than in our study for the 12 UTC runs. This may be due to the different study periods (one year versus six months). Forecasts of wind speed at 10 m were difficult to compare with forecasts at 2 m in our study, yet the same trend was observed, with ENS performing better than HRES at longer lead times. This trend was also observed for air temperature.

Zhang et al. [27] demonstrated that the errors of deep learning-based models (Graph-Cast, Pangu-Weather and Fuxi) were greater than those of HRES in extreme cold, heat and wind situations. They also found that deep learning-based models tend to underes-

timate the frequency and intensity of record-breaking events. They underestimate heat records and overestimate cold records. In our study, AIFS showed very good results compared to conventional models in forecasting the heat spike in July and the heat wave in August. However, although these days were hot by Belgian standards, they were not record-breaking. The ability of AIFS to forecast extreme hot and cold events should be evaluated over an extended period.

Moldovan et al. [28] studied the performance of HRES and AIFS SingleV1. As in our study, AIFS had a lower RMSE than HRES for temperature and radiation at all lead times.

When comparing AIFS with other models, it is important to note that the differences between these models do not lie solely in the distinction between data-driven and physics-based approaches. AIFS operates at a much coarser resolution than local models such as AROME, ICON-D2, and HARMONIE, and spatial resolution may explain part of the observed differences in model performance.

4.3. Precipitation Forecast and Its Verification

Verifying precipitation forecasts over short periods (a few hours or less) is particularly complex due to high temporal and spatial variability. Regional weather forecast models can simulate realistic heavy precipitation events, but due to the stochastic nature of convection triggering, they sometimes fail to predict these events at the right time and/or the right location. Evaluating precipitation forecasts from high spatial and temporal resolution models presents challenges when applying the same scoring systems as low-resolution models [29,30]. A degradation of these scores is generally observed for high-resolution models, even though these models appear to provide relevant information on the structure and intensity of certain weather phenomena. In the case of intense rainfall during convective thunderstorms, a model with a low spatial resolution tends to spread the maximum rainfall over a larger area, leading to lower local accumulations, rather than accurately simulating highly localised rainfall. A high-spatial-resolution model generally represents the spatial structure of the storm better, but it typically has a localisation error that causes a “double penalty” [30,31]. The first penalty arises from failing to predict the event in the correct location, and the second from the false alarm at the incorrect location. Thus, a small localisation error can have a significant negative impact on the scores of the high-resolution model. In our study, regarding quantitative precipitation forecasts, some high-resolution regional models exhibited higher errors than the global models HRES and AIFS. The double penalty also applies to timing errors in high-temporal-resolution outputs, typically hourly precipitation accumulation rather than daily accumulation. There are many alternative scoring methods for evaluating high-resolution models without subjecting them to double penalty. Small spatial errors can be identified by examining the forecasts in the grid cells surrounding the target cell, and small timing errors can be detected by checking the forecasts for the hours preceding and following the exact lead time.

In this study, the temporal errors were limited because precipitation was accumulated on a daily basis, but there could be spatial errors, and the selected metrics were subject to the double penalty. This method was chosen for its simplicity. Forecast models downloaded from open-meteo.com were stored only for grid cells containing the stations. Thus, it was not possible to analyse the spatial structure of the forecast in the neighbourhood of the point of interest. On the one hand, penalising the localisation errors may make sense from the farmer’s point of view. Farmers want to know whether rain is likely to fall on their specific fields. On the other hand, farmers often manage several plots of land, sometimes separated by dozens of kilometres. They may prefer to know that a thunderstorm is about to occur in the region (even if the exact localisation is uncertain) rather than receive a low-resolution forecast unable to detect the thunderstorm. In any case, future studies

should ensure that forecasts are stored in the neighbourhood of points of interest in order to assess spatial errors.

Another question when dealing with precipitation occurrence is the choice of the threshold to separate rainy and non-rainy days. Rossa et al. [31] reported that thresholds ranging from 0.1 to 0.3 mm were commonly used for daily accumulations. In this study, a threshold of 1 mm was selected, in agreement with Nitha et al. [32]. Another option would be to consider several rainfall thresholds, for example thresholds for light, moderate and heavy rainfall, in order to assess forecast performance for different rainfall classes [33].

Pan et al. [34] showed that AIFS estimated precipitation in China with a lower RMSE than HRES. We also found lower RMSE values for AIFS than for HRES, at all the lead times. However, the RMSE should not be the only metric considered. When precipitation occurrence was taken into account, HRES performed better than AIFS, which largely overestimated the number of precipitation days.

4.4. Use of Weather Forecasts in Decision-Support Systems for Agriculture

Part of this study focused on a decision-support tool designed to identify optimal spraying windows in agriculture. Choosing the right spraying time is crucial for farmers, as they seek to maximise product efficacy and avoid losses in the environment. The most suitable models to feed the decision-support tool were ICON-D2 and AIFS. The results of this study suggest that farmers should consult the spraying decision-support tool daily and avoid planning daytime spraying activities more than 2 days in advance. Decision-support tool designers should also be aware of the forecast limitations and avoid displaying spraying recommendations one week in advance. Although these spraying decision tools are common in agriculture, no other study (to our knowledge) has described the impact of lead time on the spraying recommendations.

Aleksovska et al. [35] calibrated ensemble weather forecasts against observations from weather stations. The same authors then detailed how to combine these post-processed ensemble weather forecasts to feed a temperature-driven pest development model for vineyards. Ensemble predictions from the regional AROME model were the most accurate but were limited to 48 h, whereas global ensemble predictions from ARPEGE or ENS could extend the forecast horizon. The authors developed a method to smoothly combine high-resolution AROME ensemble members at short range with larger-scale ensemble members at longer range. For the spraying decision-support tool presented in this paper, such a combination approach may be useful if long-range forecasts improve in the future. For now, it may be best to limit forecasts to a 48 h lead time, and avoid displaying longer lead times, as they could mislead farmers.

This recommendation to limit the use of weather forecasts to the first 48 h is valid only for the simplified spraying decision-support tool presented in this study. This recommendation could probably be extended to other spraying decision-support tools, but not to decision support tools for other agricultural applications, which may use longer-range weather forecasts. For example, Tarraf et al. [36] studied a decision-support tool for water stress in vineyards and concluded that, in most scenarios for lead times shorter than 120 h, uncertainties from model parameters and past weather observations were more important than uncertainties associated with weather forecasts. This water stress model can afford to use longer-range, less reliable weather forecasts than the spraying model because the consequences of forecast errors are less critical. Unexpected rain would force the complete cancellation of the spraying operation, whereas it would only slightly affect the vineyard's water status, which is also influenced by previous rainfall and modelled crop development.

Different errors in the input weather forecasts can lead to different consequences for the farmer and the environment. For the spraying decision-support tool, an unexpected

rain just before spraying leads to the cancellation of the operation and to a loss of time for the farmer. But an unexpected rain just after spraying may lead to environmental pollution and the need for the farmer to repeat the spraying operation to ensure the efficacy of the product and avoid crop damage due to the absence of adequate protection, thus also having economic consequences. The same forecasting error, but occurring at different times, gives rise to different risks. Spraying decision-support tools could weight the interpretation of ensemble forecasts differently depending on the risk level associated with the product's sensitivity to specific weather conditions and the planned timing of the spraying operation.

4.5. Limitations of This Study

The study period covered six months, which remains short to capture the full range of meteorological variability. It would be inappropriate to extrapolate the results to other time periods or to different regions. This study therefore provides only an indication of the performance of the evaluated weather forecast models for a specific region and a specific time frame.

It should also be noted that the models were compared only at lead times that were multiples of 6 h. Yet, many decision-support tools used in agriculture require input weather data with an hourly time step, even for longer lead times. Models with temporal resolution coarser than one hour for long lead times therefore require interpolation to feed decision-support tools. Such interpolations may prove less accurate than forecasts generated directly for every hour. In this sense, the MAR model may have an advantage over other models that provide a long forecast horizon, but at a lower temporal resolution. The results of this study may therefore lead to an underestimation of the value of the MAR model. It should also be noted that the temporal resolution of AIFS was limited to 6 h, even for short lead times. Although this model performed very well for the lead times considered in this study, its practical usefulness for agricultural decision-support systems may be limited.

Another important limitation lies in the rapid evolution of numerical weather prediction models. Forecasting systems are continuously updated through improvements in model physics, data assimilation techniques, and spatial resolution, each of which can significantly affect forecast performance. Consequently, the version of a model used at a given time may differ substantially from the version used later, potentially leading to different results. It is therefore essential to closely monitor model developments and future upgrades when interpreting forecast evaluations or implementing such models in operational decision-support tools.

This study was based on the stations of the PAMESEB network, which served as the reference for agricultural areas, but these stations differ in some respects from stations from the national weather service, the Royal Meteorological Institute of Belgium (RMI). Table A11 presents the RMSE and bias between a PAMESEB weather station and an RMI weather station.

Finally, the usefulness of the different weather forecast models for helping farmers determine the best time to spray was investigated only through the outputs of a simplified decision-support tool. It was not based on actual farmer decisions or on field measurements of the agronomic efficacy or the environmental pollution resulting from spraying. Moreover, the simplified spraying decision-support tool did not include all the criteria that may influence the decision to spray. Its limitations are detailed in Section 2.5. For example, root herbicides require the soil to be sufficiently wet, while excessive rainfall in the days following spraying should be avoided. A more advanced spraying decision-support tool could check the occurrence of rain in the days before or after the spraying date to ensure that the soil is, or will become, sufficiently wet. A weather forecast model that did not perform

best with the simplified tool might prove to be the most suitable for a more advanced tool, for example if it performs better at forecasting light rainfall.

5. Conclusions

This study evaluated the performance of eight weather forecast models for five near-surface meteorological variables and assessed their suitability for use in a simple agricultural decision-support tool designed to help farmers identify optimal spraying conditions.

The following conclusions were drawn based on data from 22 locations in Wallonia (Belgium) over a relatively dry six-month period, from March to August 2025. They are valid under these conditions and only for the versions of the weather forecast models used in this study. ICON-D2 and AIFS were the best-performing models for predicting air temperature and relative humidity up to two days ahead. Beyond two days, AIFS performed best. It was only outperformed by ENS for forecasting air temperature at ten days. ICON-D2 was the best model for predicting wind speed up to two days ahead. After two days, ENS performed best. AIFS was the best model to predict global radiation, followed by HRES. AIFS was one of the best models for quantitative and qualitative (rainy or non-rainy day) precipitation forecasting, although it tended to predict too many precipitation events. Model performance was spatially consistent across the different weather stations in Wallonia, indicating homogeneous behaviour among the evaluated forecast models within the study area.

When the forecast models were tested as inputs to a simplified agricultural decision-support tool, ICON-D2 and AIFS yielded the most reliable recommendations for spraying conditions. However, based on the results of this study, such spraying recommendations should not be based on weather variables forecast more than 48 h in advance. After this lead time, the weather forecasts led to poor spraying recommendations, especially for spraying in the daytime. It should also be noted that AIFS has a temporal resolution of 6 h. Although it ranked among the best models for the evaluated lead times, its practical usefulness for agricultural decision-support tools is limited, as these often rely on hourly weather data.

Author Contributions: Conceptualization, V.A., S.D. and D.R.; Methodology, V.A., S.D. and D.R.; Software, V.A., S.D. and J.P.H.; Validation, V.A. and S.D.; Formal Analysis, V.A. and S.D.; Investigation, V.A. and S.D.; Resources, V.M., J.P.H. and D.R.; Data Curation, V.A., S.D., V.M. and J.P.H.; Writing—Original Draft Preparation, V.A. and S.D.; Writing—Review & Editing, V.A., S.D., V.M., J.P.H., V.P. and D.R.; Visualization, V.A. and S.D.; Supervision, V.P. and D.R.; Project Administration, V.P. and D.R.; Funding Acquisition, V.P. and D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Wallonia recovery plan: project 142, Agromet II, led by the Agriculture, Natural Resources, Environment section of the Public Service of Wallonia and the Agence du Numérique (AdN).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors are grateful to Xavier Fettweis from the University of Liège and Fabian Debal from the Royal Meteorological Institute for wise advice and relevant feedback on the verification of weather forecast models. The authors also want to thank Xavier Fettweis for sharing the data from the MAR model.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AIFS	Data-driven Weather Forecast Model of the ECMWF
ECMWF	European Center for Medium-Range Weather Forecasts
ETS	Equitable Threat Score
GFS	Global Forecast System
ENS	Ensemble Weather Forecast Model of the ECMWF
FAR	False Alarm Ratio
FBI	Frequency Bias Index
HRES	Deterministic physical model of the ECMWF
MAE	Mean Absolute Error
MBE	Mean Bias Error
POD	Probability Of Detection
RMSE	Root Mean Square Error
UTC	Coordinated Universal Time

Appendix A

Table A1. Missing 12 UTC and 00 UTC runs.

Model	Missing 12 UTC Runs	Missing 00 UTC Runs
HRES	0	0
ENS	0	0
AIFS	0	0
GFS	9	9
MAR	10	14
AROME	6	45
ICON-D2	5	26
HARMONIE	26	162

Appendix B

Table A2. RMSE (°C) of air temperature forecasts for the 00 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, orange, and red correspond to a, b, c, and d, respectively. Subsequent letters are also shown in red.

Model	Forecast Lead Time (H) and Corresponding UTC Hour											
	0 00	6 06	12 12	18 18	24 00	36 12	48 00	72 00	96 00	120 00	168 00	240 00
ERA5	1.21 ^a											
AROME	1.3 ^b	1.19 ^b	1.36 ^c	1.31 ^b	1.67 ^c	1.72 ^c	1.91 ^c					
ICON-D2	1.23 ^{ab}	1.08 ^a	1.18 ^a	1.24 ^a	1.46 ^{ab}	1.45 ^{ab}	1.58 ^b					
MAR	3.24 ^d	2.01 ^d	1.72 ^d	1.66 ^d	3.24 ^e	1.84 ^d	3.23 ^e	3.00 ^e	3.30 ^d	3.49 ^d	3.87 ^d	4.06 ^e
HRES	1.31 ^b	1.07 ^a	1.37 ^c	1.46 ^c	1.53 ^b	1.52 ^b	1.61 ^b	1.80 ^c	2.00 ^b	2.21 ^b	2.85 ^b	3.66 ^c
ENS	1.27 ^b	1.03 ^a	1.30 ^{bc}	1.43 ^c	1.46 ^a	1.42 ^a	1.52 ^{ab}	1.68 ^b	1.84 ^a	2.00 ^a	2.44 ^a	2.81 ^a
GFS	1.72 ^c	1.34 ^c	2.00 ^e	1.77 ^e	2.00 ^d	2.13 ^e	2.05 ^d	2.05 ^d	2.25 ^c	2.48 ^c	3.23 ^c	4.31 ^d
AIFS	1.24 ^{ab}	1.03 ^a	1.26 ^b	1.25 ^a	1.41 ^a	1.45 ^{ab}	1.47 ^a	1.56 ^a	1.82 ^a	1.94 ^a	2.48 ^a	3.52 ^b

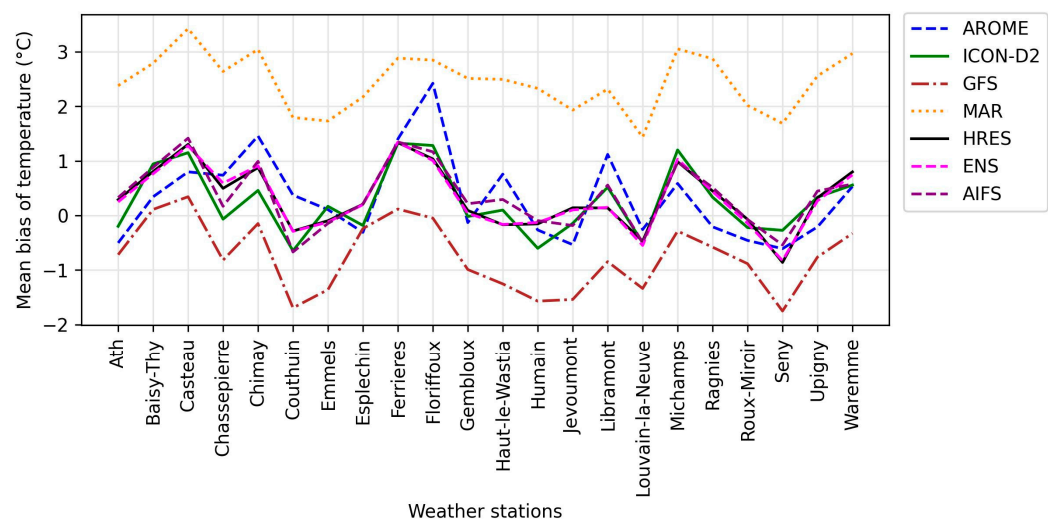


Figure A1. MBE (°C) of air temperature forecasts by weather stations for the 00 UTC runs at 24 h lead time (00:00).

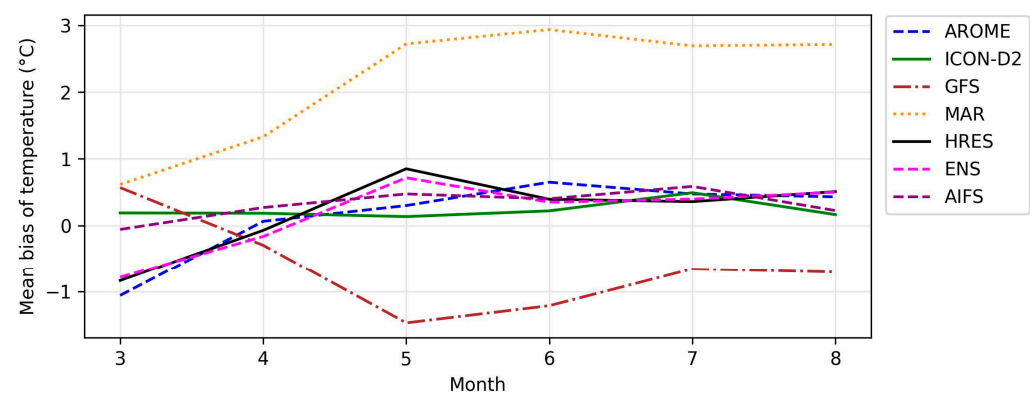


Figure A2. Monthly MBE (°C) of air temperature forecasts for the 00 UTC runs at 24 h lead time.

Appendix C

Table A3. RMSE (%) of relative humidity forecasts for the 00 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, orange, and red correspond to a, b, c, and d, respectively. Subsequent letters are also shown in red.

Model	Forecast Lead Time (H) and Corresponding UTC Hour											
	0 00	6 06	12 12	18 18	24 00	36 12	48 00	72 00	96 00	120 00	168 00	240 00
ERA5	6.9 ^a											
AROME	8.5 ^c	7.2 ^c	8.3 ^b	10.5 ^c	12.2 ^c	11.4 ^c	13.7 ^d					
ICON-D2	7.1 ^a	6.6 ^b	7.6 ^a	9.1 ^b	8.9 ^a	9.4 ^{ab}	9.3 ^a					
MAR	13.8 ^e	10.0 ^d	16.5 ^d	15.7 ^f	13.9 ^d	16.7 ^e	14.0 ^d	14.1 ^d	15.5 ^d	16.1 ^d	17.8 ^d	19.9 ^d
HRES	7.9 ^b	6.7 ^b	8.3 ^b	11.1 ^d	9.3 ^b	9.8 ^b	9.9 ^b	11.0 ^b	11.8 ^b	13.3 ^b	14.2 ^b	17.3 ^b
GFS	10.6 ^d	10.6 ^e	12.4 ^c	14.3 ^e	12.6 ^c	13.4 ^d	12.8 ^c	12.9 ^c	13.8 ^c	14.6 ^c	16.2 ^c	18.2 ^c
AIFS	7.0 ^a	6.1 ^a	8.0 ^b	8.5 ^a	9.2 ^{ab}	9.2 ^a	9.1 ^a	9.1 ^a	9.7 ^a	10.4 ^a	11.7 ^a	13.9 ^a

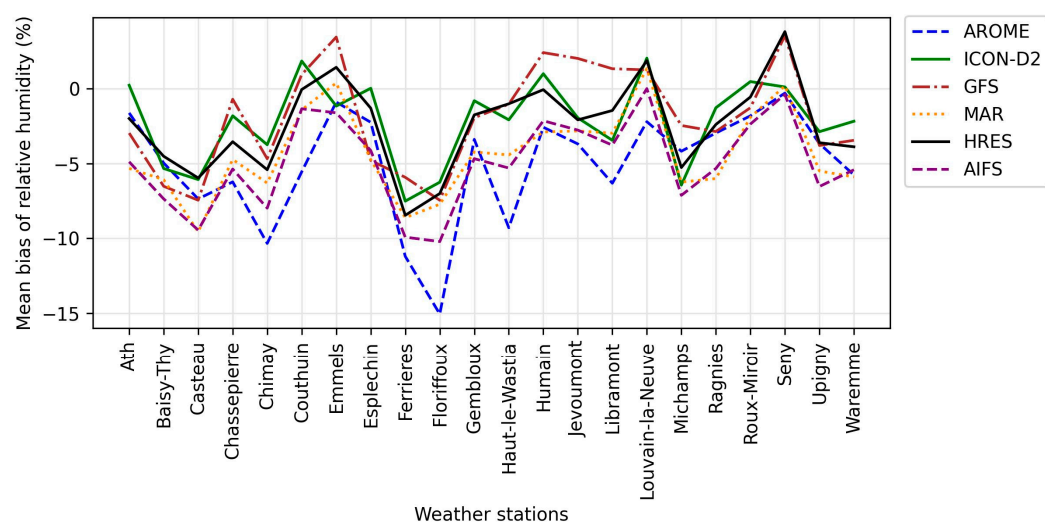


Figure A3. MBE (%) of relative air humidity forecasts by weather stations for the 00 UTC runs at 24 h lead time.

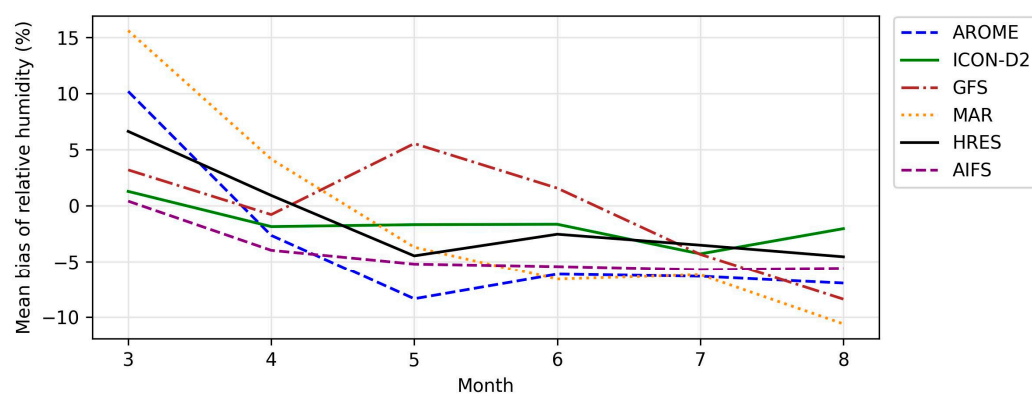


Figure A4. Monthly MBE (%) of relative humidity forecasts for the 00 UTC runs at 24 h lead time.

Appendix D

Table A4. RMSE (m s^{-1}) of wind speed forecasts for the 00 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, orange, and red correspond to a, b, c, and d, respectively. Subsequent letters are also shown in red.

Model	Forecast Lead Time (h) and Corresponding UTC Hour											
	0 00	6 06	12 12	18 18	24 00	36 12	48 00	72 00	96 00	120 00	168 00	240 00
ERA5	1.01 ^d											
AROME	0.90 ^{bc}	0.88 ^{ab}	1.30 ^b	1.04 ^{ab}	0.96 ^{bc}	1.26 ^{ab}	0.96 ^{bc}					
ICON-D2	0.75 ^a	0.84 ^a	1.27 ^{ab}	1.02 ^a	0.80 ^a	1.24 ^{ab}	0.79 ^a					
MAR	1.29 ^e	1.21 ^c	1.24 ^a	1.38 ^c	1.32 ^d	1.27 ^{ab}	1.36 ^d	1.42 ^c	1.45 ^c	1.63 ^d	1.67 ^d	1.69 ^d
HRES	0.93 ^c	0.92 ^b	1.33 ^b	1.08 ^b	0.95 ^{bc}	1.29 ^b	0.94 ^{bc}	0.99 ^b	1.06 ^b	1.16 ^c	1.32 ^c	1.27 ^b
ENS	0.93 ^c	0.91 ^b	1.28 ^{ab}	1.02 ^a	0.93 ^b	1.22 ^a	0.93 ^b	0.96 ^a	1.00 ^a	1.03 ^a	1.07 ^a	1.09 ^a
AIFS	0.89 ^b	0.93 ^b	1.30 ^b	1.04 ^{ab}	0.98 ^c	1.27 ^{ab}	0.97 ^c	1.00 ^b	1.04 ^b	1.10 ^b	1.19 ^b	1.39 ^c

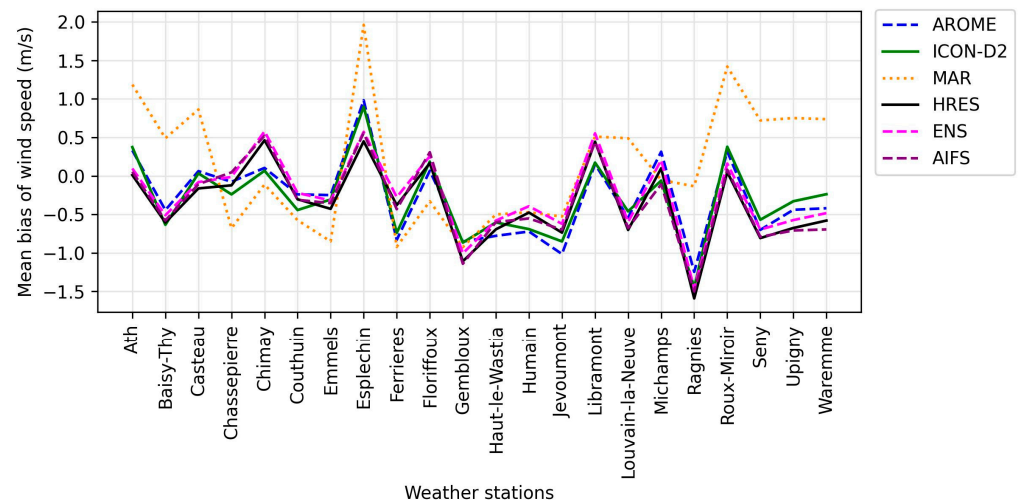


Figure A5. MBE (m s^{-1}) of wind speed forecasts by weather stations for the 00 UTC runs at 36 h lead time.

Appendix E

Table A5. RMSE (mm) of daily precipitation forecasts (≥ 0.1 mm) for the 00 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, and orange correspond to a, b, and c, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour						
	24 00	48 00	72 00	96 00	120 00	168 00	240 00
AROME	5.4 ^b	6.2 ^c					
ICON-D2	5.7 ^b	5.6 ^{bc}					
MAR	4.8 ^{ab}	4.9 ^b	4.6 ^b	5.0 ^b	5.9 ^b	5.6 ^b	6.6 ^b
HRES	4.8 ^{ab}	4.8 ^{ab}	4.7 ^b	5.4 ^b	5.8 ^b	6.6 ^c	7.4 ^c
AIFS	4.5 ^a	4.1 ^a	3.9 ^a	4.1 ^a	4.7 ^a	4.9 ^a	5.9 ^a

Table A6. ETS of rainy day (daily precipitation ≥ 1 mm) for the 00 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, and orange correspond to a, b, and c, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour						
	24 00	48 00	72 00	96 00	120 00	168 00	240 00
AROME	0.62 ^a	0.45 ^b					
ICON-D2	0.59 ^a	0.45 ^b					
MAR	0.38 ^c	0.34 ^c	0.32 ^b	0.27 ^c	0.27 ^b	0.17 ^b	0.08 ^c
HRES	0.51 ^b	0.51 ^a	0.42 ^a	0.35 ^b	0.3 ^{ab}	0.22 ^a	0.13 ^b
AIFS	0.53 ^b	0.5 ^{ab}	0.42 ^a	0.39 ^a	0.32 ^a	0.22 ^a	0.18 ^a

Table A7. FAR (%) of rainy day (daily precipitation ≥ 1 mm) for the 00 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, and orange correspond to a, b, and c, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour						
	24 00	48 00	72 00	96 00	120 00	168 00	240 00
AROME	25 ^a	36 ^{ab}					
ICON-D2	24 ^a	32 ^a					
MAR	46 ^c	50 ^c	51 ^b	56 ^b	55 ^b	63 ^b	69 ^b
HRES	30 ^b	34 ^{ab}	41 ^a	48 ^a	48 ^a	56 ^a	61 ^a
AIFS	31 ^b	38 ^b	45 ^a	48 ^a	51 ^{ab}	60 ^{ab}	60 ^a

Table A8. POD (%) of rainy day (daily precipitation ≥ 1 mm) for the 00 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, orange, and red correspond to a, b, c, and d, respectively. Subsequent letters are also shown in red.

Model	Forecast Lead Time (H) and Corresponding UTC Hour						
	24 00	48 00	72 00	96 00	120 00	168 00	240 00
AROME	91 ^a	76 ^d					
ICON-D2	85 ^b	70 ^e					
MAR	89 ^{ab}	83 ^c	78 ^b	76 ^b	77 ^b	59 ^b	50 ^c
HRES	85 ^b	89 ^b	81 ^b	79 ^b	69 ^c	63 ^b	54 ^b
AIFS	88 ^{ab}	95 ^a	93 ^a	93 ^a	86 ^a	76 ^a	77 ^a

Table A9. FBI of rainy day (daily precipitation ≥ 1 mm) for the 00 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, orange, and red correspond to a, b, c, and d, respectively. Subsequent letters are also shown in red.

Model	Forecast Lead Time (h) and Corresponding UTC Hour						
	24 00	48 00	72 00	96 00	120 00	168 00	240 00
AROME	1.22 ^b	1.19 ^b					
ICON-D2	1.12 ^a	1.03 ^a					
MAR	1.63 ^c	1.68 ^e	1.59 ^b	1.74 ^b	1.69 ^b	1.60 ^b	1.61 ^b
HRES	1.22 ^b	1.35 ^c	1.37 ^a	1.51 ^a	1.33 ^a	1.43 ^a	1.41 ^a
AIFS	1.28 ^b	1.52 ^d	1.69 ^b	1.79 ^b	1.76 ^b	1.89 ^c	1.90 ^c

Appendix F

Table A10. RMSE (W m^{-2}) of mean daily global radiation for the 00 UTC runs. For each lead time, the letters indicate statistically distinguishable groups based on the bootstrap comparison. Green, yellow, orange, and red correspond to a, b, c, and d, respectively.

Model	Forecast Lead Time (h) and Corresponding UTC Hour						
	24 12	48 12	72 12	96 12	120 12	168 12	240 12
ERA5	32.0 ^c						
MAR	40.1 ^d	42.9 ^c	49.2 ^c	54.4 ^c	65.9 ^c	77.1 ^b	88.6 ^c
HRES	29.2 ^b	32.8 ^b	39.0 ^b	41.0 ^b	50.4 ^b	59.3 ^a	78.4 ^b
AIFS	28.0 ^a	30.8 ^a	34.5 ^a	39.3 ^a	46.5 ^a	59.4 ^a	69.8 ^a

Appendix G

Table A11. Quantitative comparison between PAMESEB and RMI weather stations. Positive bias indicates an overestimation of the PAMESEB station and negative bias indicates an underestimation.

Weather Variables	RMSE	Bias
Air temperature 2 m	0.18 °C	0.10 °C
Relative humidity 2 m	3.8%	2.4%
Wind speed 2 m	0.17 m s ⁻¹	0.0 m s ⁻¹
Global radiation	24 W m ⁻²	18 W m ⁻² , even greater during summer
Precipitation	/	Underestimation on long-term accumulation

References

- Bauer, P.; Thorpe, A.; Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **2015**, *525*, 47–55. [\[CrossRef\]](#) [\[PubMed\]](#)
- Haiden, T.; Janousek, M.; Vitart, F.; Ben Bouallègue, Z.; Ferranti, L.; Prates, F.; Richardson, D. *Evaluation of ECMWF Forecasts, Including the 2021 Upgrade*; ECMWF: Reading, UK, 2022.
- Haiden, T.; Rodwell, M.J.; Richardson, D.S.; Okagaki, A.; Robinson, T.; Hewson, T. Intercomparison of Global Model Precipitation Forecast Skill in 2010/11 Using the SEEPS Score. *Mon. Weather Rev.* **2012**, *140*, 2720–2733. [\[CrossRef\]](#)
- Magnaldo, M.-A.; Libois, Q.; Riette, S.; Lac, C. Evaluation of surface shortwave downward radiation forecasts by the numerical weather prediction model AROME. *Geosci. Model Dev.* **2024**, *17*, 1091–1109. [\[CrossRef\]](#)
- Hagelin, S.; Auger, L.; Brovelli, P.; Dupont, O. Nowcasting with the AROME Model: First Results from the High-Resolution AROME Airport. *Weather Forecast.* **2014**, *29*, 773–787. [\[CrossRef\]](#)
- Häuser, B. Analysis and Improvement of the Quality of Global Wind Forecasts. Master's Thesis, Albert-Ludwigs-University Freiburg, Freiburg im Breisgau, Germany, 2021.
- Reiß, A. Global Verification of Incoming Shortwave Radiation for Several Weather Forecast Models and Different Model Combinations. Bachelor's Thesis, Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau, Germany, 2021.
- Fessler, E. Accuracy and Variability of Six Global Temperature Model Forecasts in 2018. Bachelor's Thesis, Universität Basel, Basel, Switzerland, 2019.
- Chen, K.; Han, T.; Ling, F.; Gong, J.; Bai, L.; Wang, X.; Luo, J.-J.; Fei, B.; Zhang, W.; Chen, X.; et al. The operational medium-range deterministic weather forecasting can be extended beyond a 10-day lead time. *Commun. Earth Environ.* **2025**, *6*, 518. [\[CrossRef\]](#)
- Waqas, M.; Humphries, U.W.; Chueasa, B.; Wangwongchai, A. Artificial intelligence and numerical weather prediction models: A technical survey. *Nat. Hazards Res.* **2025**, *5*, 306–320. [\[CrossRef\]](#)
- Schultz, M.G.; Betancourt, C.; Gong, B.; Kleinert, F.; Langguth, M.; Leufen, L.H.; Mozaffari, A.; Stadler, S. Can deep learning beat numerical weather prediction? *Philos. Trans. Ser. A Math. Phys. Eng. Sci.* **2021**, *379*, 20200097. [\[CrossRef\]](#)
- Olivetti, L.; Messori, G. Advances and prospects of deep learning for medium-range extreme weather forecasting. *Geosci. Model Dev.* **2024**, *17*, 2347–2358. [\[CrossRef\]](#)
- Lang, S.; Alexe, M.; Chantry, M.; Dramsch, J.; Pinault, F.; Raoult, B.; Clare, M.C.A.; Lessig, C.; Maier-Gerber, M.; Magnusson, L.; et al. AIFS—ECMWF's data-driven forecasting system. *arXiv* **2024**, arXiv:2406.01465. [\[CrossRef\]](#)
- Bodnar, C.; Bruinsma, W.P.; Lucic, A.; Stanley, M.; Allen, A.; Brandstetter, J.; Garvan, P.; Riechert, M.; Weyn, J.A.; Dong, H.; et al. A foundation model for the Earth system. *Nature* **2025**, *641*, 1180–1187. [\[CrossRef\]](#) [\[PubMed\]](#)
- Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **2023**, *619*, 533–538. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lam, R.; Sanchez-Gonzalez, A.; Willson, M.; Wirsberger, P.; Fortunato, M.; Alet, F.; Ravuri, S.; Ewalds, T.; Eaton-Rosen, Z.; Hu, W.; et al. Learning skillful medium-range global weather forecasting. *Science* **2023**, *382*, 1416–1421. [\[CrossRef\]](#) [\[PubMed\]](#)
- Dandriofosse, S.; Jago, A.; Huart, J.P.; Michaud, V.; Planchon, V.; Rosillon, D. Automatic quality control of weather data for timely decisions in agriculture. *Smart Agric. Technol.* **2024**, *8*, 100445. [\[CrossRef\]](#)
- Seity, Y.; Brousseau, P.; Malardel, S.; Hello, G.; Bénard, P.; Bouttier, F.; Lac, C.; Masson, V. The AROME-France Convective-Scale Operational Model. *Mon. Weather Rev.* **2011**, *139*, 976–991. [\[CrossRef\]](#)
- Reinert, D.; Prill, F.; Frank, H.; Denhard, M.; Baldauf, M.; Schraff, C.; Gebhardt, C.; Marsigli, C.; Förstner, J.; Zängl, G.; et al. *DWD Database Reference for the Global and Regional ICON and ICON-EPS Forecasting System—Version 2.5.4*; Deutscher Wetterdienst: Offenbach, Germany, 2026.
- Bengtsson, L.; Andrae, U.; Aspelien, T.; Batrak, Y.; Calvo, J.; Rooy Wde Gleeson, E.; Hansen-Sass, B.; Homleid, M.; Hortal, M.; Ivarsson, K.-I.; et al. The HARMONIE-AROME Model Configuration in the ALADIN-HIRLAM NWP System. *Mon. Weather Rev.* **2017**, *145*, 1919–1935. [\[CrossRef\]](#)
- de Rooy, W.C.; Siebesma, P.; Baas, P.; Lenderink, G.; de Roode, S.R.; de Vries, H.; van Meijgaard, E.; Meirink, J.F.; Tijn, S.; van 't Veen, B. Model development in practice: A comprehensive update to the boundary layer schemes in HARMONIE-AROME cycle 40. *Geosci. Model Dev.* **2022**, *15*, 1513–1543. [\[CrossRef\]](#)
- Grailet, J.-F.; Hogan, R.J.; Ghilain, N.; Bolsée, D.; Fettweis, X.; Grégoire, M. Inclusion of the ECMWF ecRad radiation scheme (v1.5.0) in the MAR (v3.14), regional evaluation for Belgium, and assessment of surface shortwave spectral fluxes at Uccle. *Geosci. Model Dev.* **2025**, *18*, 1965–1988. [\[CrossRef\]](#)
- 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. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 1999–2049. [\[CrossRef\]](#)
- Zippenfenig, P. *Open-Meteo.com Weather API*, version 1.4.0 (Computer Software); Zenodo: Geneva, Switzerland, 2024. [\[CrossRef\]](#)
- Goff, J.A. Saturation pressure of water on the new Kelvin scale. *Amer. Soc. Heat. Vent. Eng. trans.* **1957**, *63*, 347–354.
- Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213–220. [\[CrossRef\]](#) [\[PubMed\]](#)

27. Zhang, Z.; Fischer, E.; Zscheischler, J.; Engelke, S. Numerical models outperform AI weather forecasts of record-breaking extremes. *arXiv* **2025**, arXiv:2508.15724. [[CrossRef](#)]
28. Moldovan, G.; Pinnington, E.; Nemesio, A.P.; Lang, S.; Bouallègue, Z.B.; Dramsch, J.; Alexe, M.; Cruz, M.S.; Hahner, S.; Cook, H.; et al. An update to ECMWF's machine-learned weather forecast model AIFS. *arXiv* **2025**, arXiv:2509.18994. [[CrossRef](#)]
29. Zimmer, M.; Wernli, H. Verification of quantitative precipitation forecasts on short time-scales: A fuzzy approach to handle timing errors with SAL. *Meteorol. Z.* **2011**, *20*, 95–105. [[CrossRef](#)] [[PubMed](#)]
30. Stein, J.; Pailleux, J.; Stoop, F.; Amodei, M.; Dupont, O.; Mayoka, M.; Pouponneau, F.; Sanchez, I. La vérification des prévisions météorologiques à Météo-France. *Météorologie* **2015**, *8*, 40. [[CrossRef](#)]
31. Rossa, A.; Nurmi, P.; Ebert, E. Overview of methods for the verification of quantitative precipitation forecasts. In *Precipitation: Advances in Measurement, Estimation and Prediction*; Michaelides, S., Ed.; Springer: Berlin/Heidelberg, German, 2008; pp. 419–452. [[CrossRef](#)]
32. Nitha, V.; Pramada, S.K.; Praseed, N.S.; Sridhar, V. Performance Evaluation of Numerical Weather Prediction Models in Forecasting Rainfall Events in Kerala, India. *Atmosphere* **2025**, *16*, 372. [[CrossRef](#)]
33. Le Duc Saito, K.; Seko, H. Spatial-temporal fractions verification for high-resolution ensemble forecasts. *Tellus A Dyn. Meteorol. Oceanogr.* **2013**, *65*, 18171. [[CrossRef](#)]
34. Pan, L.; Zhang, H.; Liang, M.; Li, P.; Cao, C.; Du, L.; Liu, J. Comparative Analysis of Precipitation Forecasts between the ECMWF Artificial Intelligence System (ECMWF AIFS) and Its Integrated Forecast System (IFS). *Weather Forecast.* **2025**, *40*, 1653–1670. [[CrossRef](#)]
35. Aleksovska, I.; Raynaud, L.; Faivre, R.; Brun, F.; Raynal, M. Design and Evaluation of Calibrated and Seamless Ensemble Weather Forecasts for Crop Protection Applications. *Weather Forecast.* **2021**, *36*, 1329–1342. [[CrossRef](#)]
36. Tarraf, B.; Brun, F.; Aldebert, K.; Debord, C.; Raynaud, L.; Roux, S. Comparing the impact of weather forecasts, observation and parameterization uncertainties when predicting water stress in vineyards. *Comput. Electron. Agric.* **2025**, *238*, 110825. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.