ECMWF Feature article

from Newsletter Number 185 – Autumn 2025

EARTH SYSTEM SCIENCE

AIFS ENS becomes operational

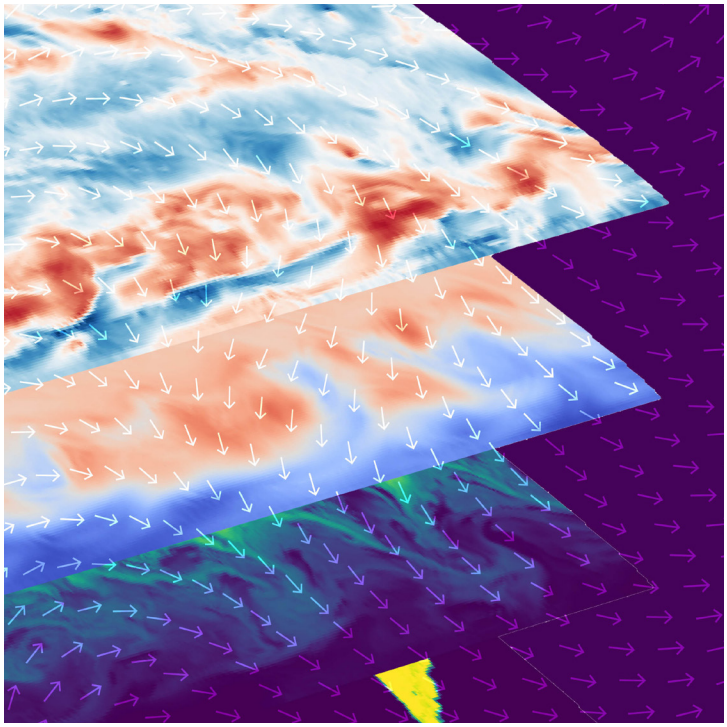

Illustration of the Anemoi ERA5 dataset

www.ecmwf.int/en/about/media-centre/media-resources

doi: 10.21957/hg1z-pe65

This article appeared in the Earth system science section of ECMWF Newsletter No. 185 – Autumn 2025, pp. 20–24

AIFS ENS becomes operational

Simon Lang, Linus Magnusson

In recent years, machine-learned (ML) weather forecasting systems such as FourCastNet, Pangu-Weather, GraphCast and the Artificial Intelligence Forecasting System (AIFS) have shown that they can produce highly skilful deterministic predictions that match or exceed those from physics-based models. These forecast models learn how to forecast by advancing the atmospheric state from one analysis date to the next during training. When in forecast mode, they run autoregressively – they evolve the atmospheric state forward in time from their own predictions.

However, the first-generation ML models typically tend to produce overly smooth fields, lacking the small-scale features that are observed in reality. This is because they are trained using a mean-squared error (MSE) objective, which can lead to reduced variability in the forecast fields.

Smoothing reduces differences among ensemble members, which decreases the spread, measured by the ensemble standard deviation, and can make the forecast unreliable. In other words, the forecasted probabilities do not match reality.

Probabilistically trained models have now been developed that address these limitations. For example, denoising diffusion-based models that start with random noise and gradually refine it, like GenCast and AIFS-Diffusion, and models trained with a loss based on a proper score, which encourages realistic atmospheric variability of forecast fields, like AIFS-CRPS. Compared to diffusion-based models, the direct optimisation of proper scores such as the Continuous Ranked Probability Score (CRPS) provides several practical advantages. First, the model can learn how to forecast across many steps. Second, generating ensemble members requires just one model evaluation per forecast step. This makes them potentially orders of magnitude more computationally efficient to run than diffusion-based models that rely on a sampler for forecasting, requiring the model to be called many times for each single forecast step. Tests have also shown that AIFS-CRPS gives more accurate forecasts than diffusion-based models. Due to the advantages demonstrated by AIFS-CRPS, the CRPS-based training approach has been adopted by others, including FourCastNet 3 by NVIDIA and Functional Generative Networks (FGN) by Google DeepMind.

After a testing phase to assess real-time forecast performance, we have now transitioned AIFS-CRPS into an operational system. It is named AIFS ENS to distinguish it from the deterministically trained AIFS Single.

Model description

AIFS ENS has been trained on 38 years of ERA5 reanalysis data (1979–2017) and eight years of data from the Integrated Forecasting System (IFS) operational analyses (2016–2023). Like the AIFS Single, the model uses an encoder-processor-decoder architecture in which the encoder reduces the input to a lower-resolution internal grid, the data go through the processor, and the decoder projects it back to the

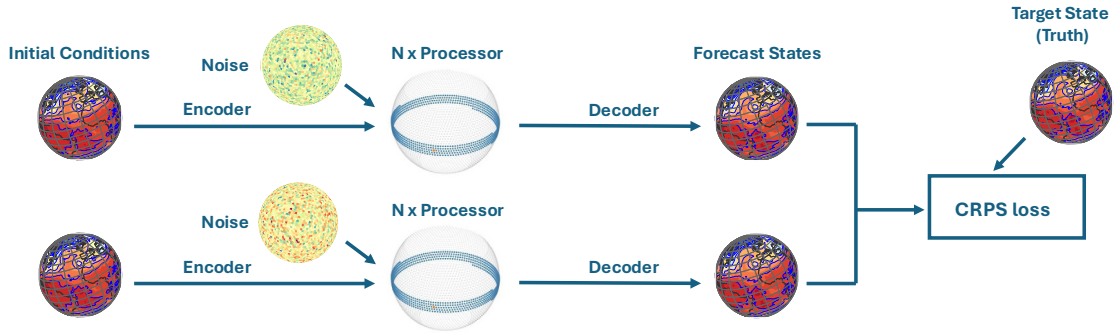

FIGURE 1 AIFS ENS training, showing the encoder-processor-decoder architecture in which the encoder reduces the input to a lower-resolution internal grid, the data are processed, and the decoder projects it back to the output grid. The loss is computed from two ensemble members in training.

 doi: 10.21957/hg1z-pe65

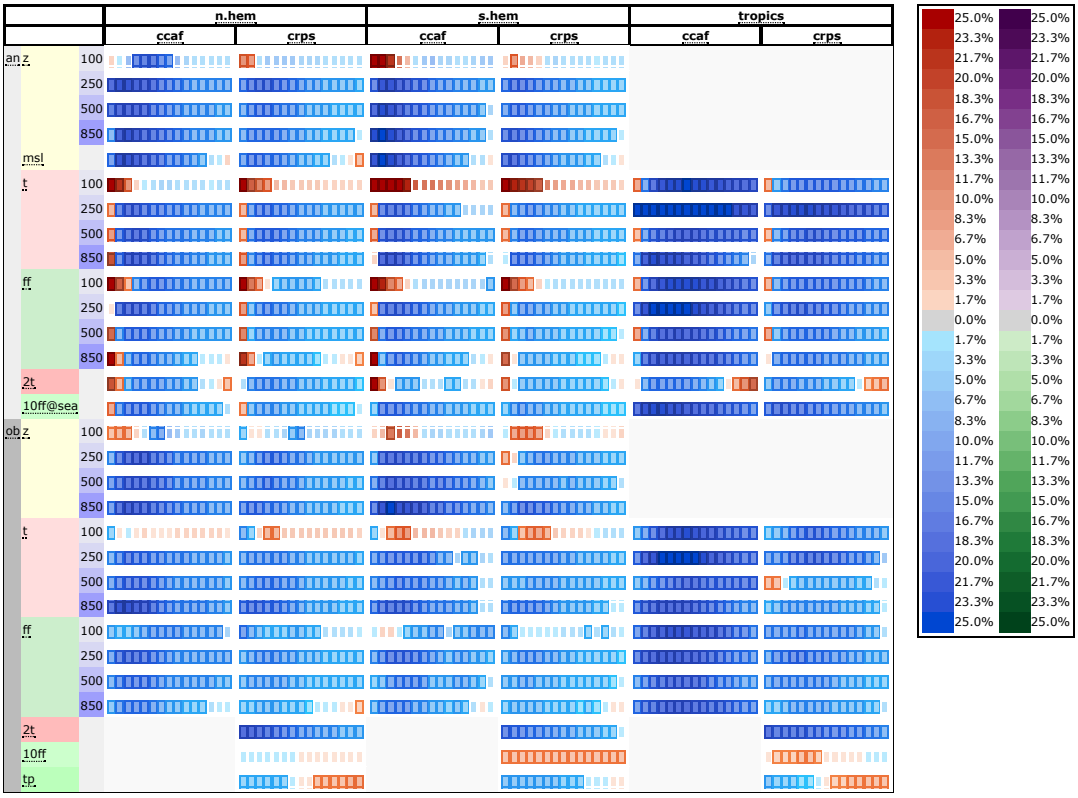

FIGURE 2 Scorecard comparing AIFS ENS with IFS ENS. Blue shading indicates where the AIFS ENS outperforms the IFS ENS and red shading indicates a degradation in forecasting skill.

output grid (Figure 1). The encoder and decoder use transformer-based graph neural networks, while the processor is a transformer with a sliding attention window, which processes information across a latitudinal band. AIFS ENS has a total of 229 million parameters and a spatial resolution of approximately 30 km. Ensemble members are generated by introducing noise within the processor. When generating forecasts, as many members as required can be generated by providing different random seeds to the model. Two ensemble members are used in training.

AIFS ENS currently has 50 perturbed members and one control member. However, unlike the physics-based IFS ENS, where the control member uses unperturbed initial conditions and the forecast model is unperturbed as well, the AIFS ENS control member also starts from unperturbed initial conditions but still includes model-generated variations during the forecast. This means that all AIFS ENS members, including the control, are stochastic forecasts that represent different outcomes from a learned forecast distribution.

Evaluation of forecast skill

AIFS ENS forecasts have been evaluated against IFS analyses and against radiosonde and SYNOP observations. Figure 2 displays the resulting scorecard, with blue shading indicating where AIFS ENS outperforms the IFS ENS and red shading indicating a degradation in forecasting skill. This scorecard shows forecast improvements reach up to 25% and that AIFS ENS has higher forecast skill for upper-air variables. The skill improvements result from reductions in both bias and random component forecast errors. Degradations, however, are seen for forecasts of conditions higher up in the atmosphere. For example, temperature at 100 hPa shows a degradation when verified against IFS operational analyses and to a lesser degree when verified against radiosonde observations. For early lead times, AIFS ENS forecasts can appear less skilful than IFS ENS forecasts when verified against IFS analyses; however, this degradation of AIFS ENS compared to IFS ENS is not visible when SYNOP and radiosonde observations are used for verification. In terms of surface variables verified against SYNOP observations, AIFS ENS shows improved scores for 2-metre temperature for all lead times, and during the first half of the forecast for 24-hour accumulated total precipitation. For 10-metre windspeed, IFS ENS is more skilful than AIFS ENS when verified against SYNOP observations.

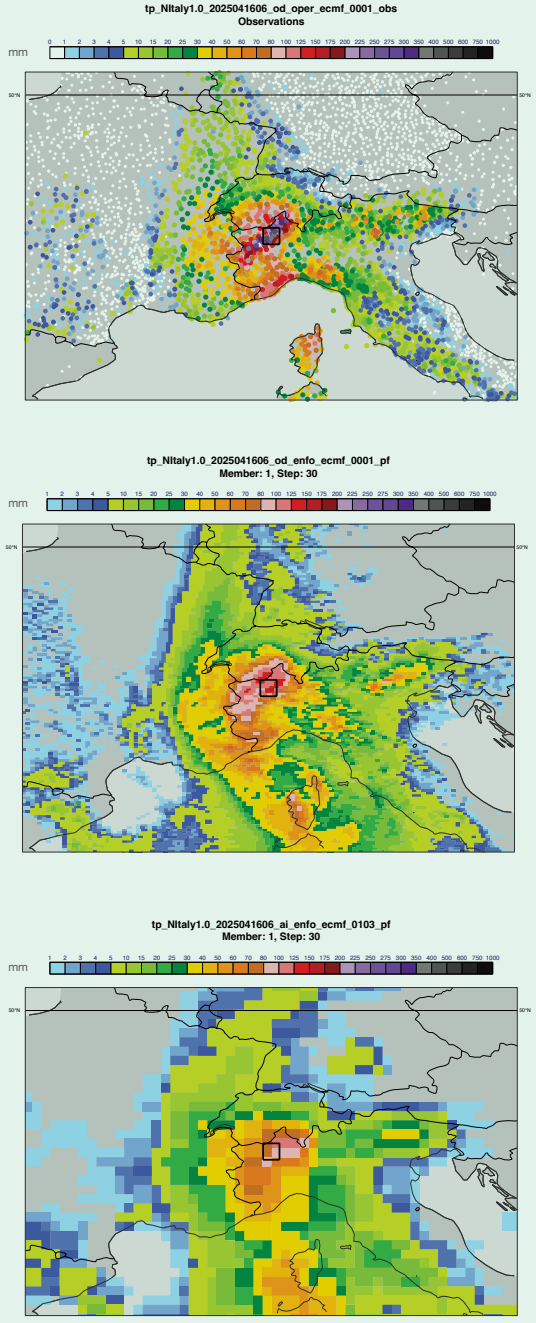

FIGURE 3 24-hour accumulated precipitation, 16 April 06 UTC – 17 April 06 UTC. Observations (top), IFS ENS member 1 (middle) and AIFS ENS member 1 (bottom).

Case study – Storm Hans in the Alps

In April 2025, Storm Hans brought extreme conditions to the southern Alps. High precipitation values, of as much as 446 mm in a 24-hour period, were recorded in northern Italy and southern Switzerland on 16 and 17 April. Figure 3 shows the 24-hour accumulated precipitation values from observations, IFS ENS member 1 and AIFS ENS member 1, of a 54-hour forecast. The perturbed members of both ensemble systems reproduce the overall precipitation pattern, though they underestimate maximum values. The underprediction of AIFS ENS seems more pronounced than in the case of IFS ENS. One likely factor is the resolution of the forecasting systems – approximately 9 km for the IFS compared to approximately 30 km for the AIFS.

Figure 4 shows the time evolution of the predicted 24-hour precipitation values from AIFS ENS and IFS ENS averaged inside the box shown in Figure 3. Both ensemble systems show a similar forecast evolution, again with AIFS ENS more strongly underpredicting the observed values. Nevertheless, both AIFS ENS and IFS ENS gave early indication of the possibility of an extreme event.

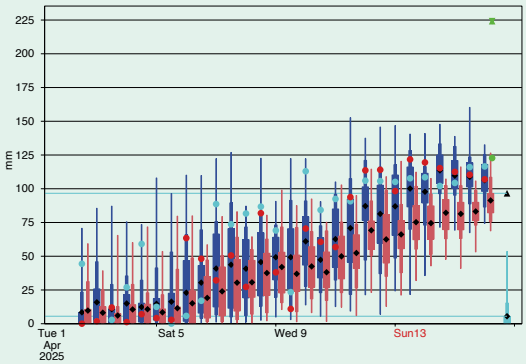

FIGURE 4 24-hour precipitation 16 April 06 UTC – 17 April 06 UTC in a 0.5-degree box in the Italian Alps (see Figure 3). Showing observation (green hourglass), Analysis (green dot), IFS ENS (blue), AIFS ENS (pink), IFS model climate (cyan).

doi: 10.21957/hg1z-pe65

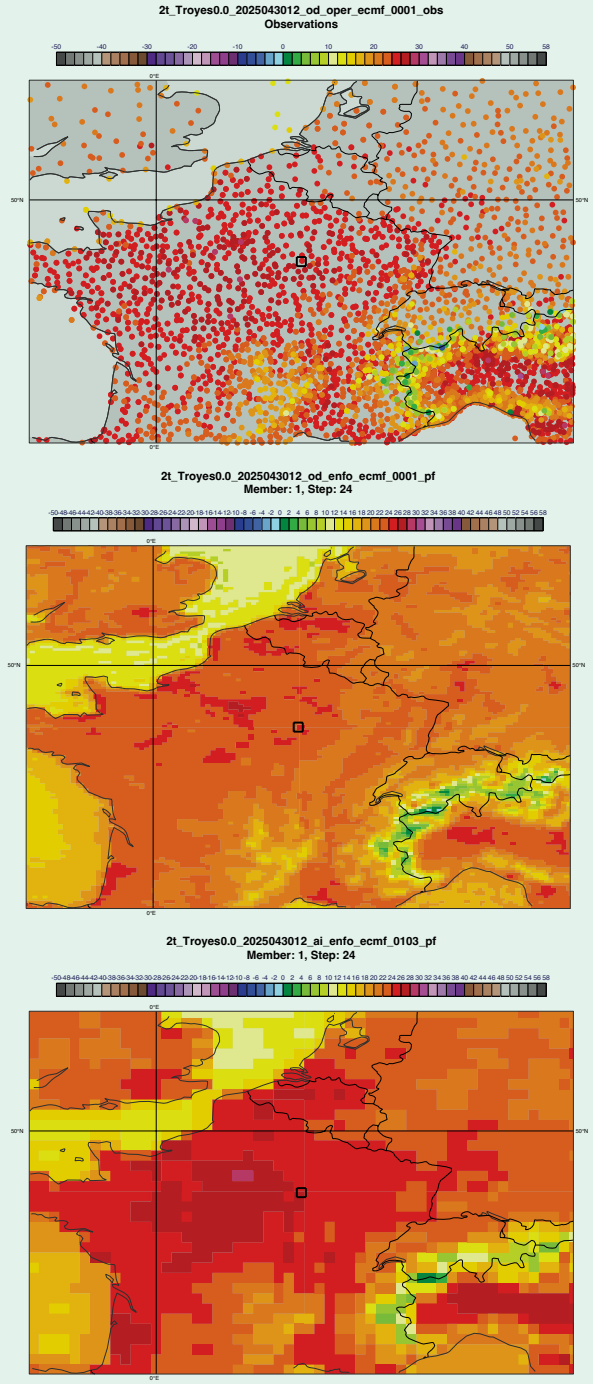

FIGURE 5 Two-metre temperature 30 April 12 UTC, observations (top), 24-hour forecast of IFS ENS member 1 (middle) and AIFS ENS member 1 (bottom).

Case study – heatwave in France

On 30 April 2025, France was struck by a heatwave. Figure 5 shows observations and short-range forecasts from IFS ENS and AIFS ENS member 1, respectively. Even for this short lead time, IFS ENS underestimated the observed temperatures. AIFS ENS, however, predicts temperature values closer to those observed.

The evolution of these forecasts is shown in Figure 6. Initially, both ensembles perform similarly; however, closer to the event, around the 7-day lead time, AIFS ENS starts to predict higher temperature values than IFS ENS. It is being investigated why the IFS ENS underestimated this heatwave and other similar conditions during late spring to early summer 2025.

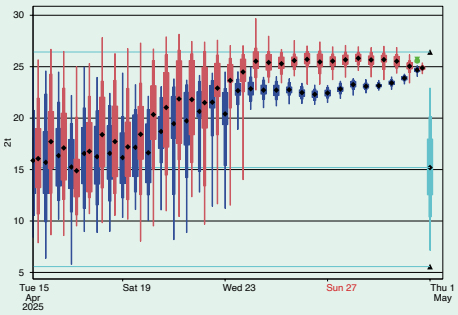

FIGURE 6 Evolution of forecasts for 2-metre temperature on 30 April 2025 at 12 UTC in Troyes, France. Symbols show observation (green hourglass), analysis (green dot), IFS ENS (blue), AIFS ENS (pink) and IFS model climate (cyan).

Limitations

While AIFS ENS shows strong forecast performance, as measured by scores for upper-air and surface conditions, some limitations remain. For example, anomalous values in mean sea level pressure and low-level temperature fields can develop, especially in regions with complex topography such as mountain ranges. Another example is 2-metre temperature forecasts, which can show degradations beyond day seven, with unrealistic cold biases appearing episodically in high-altitude locations and specific regions, including the Red Sea.

doi: 10.21957/hg1z-pe65

The model also exhibits systematic biases in precipitation forecasting, generating spurious trace amounts of rainfall or snowfall (typically less than 0.1 mm per 6-hour period) in arid regions. This issue becomes particularly apparent when examining longer accumulation periods.

Cloud cover forecast, while skilful, displays a noticeably coarser spatial resolution compared with other atmospheric variables, resulting in a blocky appearance. This limitation affects the model's ability to represent fine-scale cloud structures that are important for understanding local weather phenomena.

Finally, AIFS ENS is currently overdispersive for a range of upper-air variables – in other words, the ensemble standard deviation is larger than the root mean squared error of the ensemble mean. It is likely that this is related to the fact that AIFS ENS uses the same initial perturbation as the IFS ensemble. To improve the reliability of the IFS ensemble, singular vector perturbations are added to the initial conditions. Because AIFS ENS forecast errors are smaller, it might require a different initial perturbation amplitude.

The limitations described are under investigation, with improvements expected in future model cycles.

Conclusion and outlook

AIFS ENS has successfully transitioned from research to operational status, providing highly skilful ensemble forecasts. The CRPS-based training approach enables the generation of probabilistic forecasts with realistic atmospheric variability, suitable for ensemble forecasting. At the same time, models trained this way are computationally efficient, with each ensemble member requiring only one model evaluation per forecast step.

Performance evaluations comparing AIFS ENS to IFS ENS show improvements of up to 25% for upper-air variables and positive impacts for surface variables such as 2-metre temperature and total precipitation.

There remain several limitations to the model, including artefacts in mean sea level pressure, low effective resolution in cloud fields and spurious patterns in accumulated precipitation. Ongoing work aims to address these issues in future model cycles.

The operational implementation of the AIFS ensemble represents a milestone in ECMWF's machine learning journey. Further development will also focus on increasing horizontal and temporal resolution, and incorporating additional Earth system components.

Further reading:

Lang, S., M. Alexe, M.C.A. Clare, C. Roberts, R. Adewoyin, Z.B. Bouallègue et al., 2024: AIFS-CRPS: Ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. arXiv preprint arXiv:2412.15832 *https://doi.org/10.48550/arXiv.2412.15832*

European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, UK

 doi: 10.21957/hg1z-pe65