D3.5: Preliminary WeatherGenerator evaluation results

Due date of deliverable	31 July 2026
Submission date	30 July 2026
File Name	WeatherGenerator-D3.5-V1.2.docx
Work Package /Task	WP3 / Evaluation and interpretation
Organisation Responsible of Deliverable	UKMO (MetSwiss, eScience)
Author name(s)	Duncan Ackerley, Vitus Benson, Sebastian Buschow, Emilie Byermoen, Anna-Louise Ellis, Sebastian Hoffmann, Matthias Karlbauer, Ilaria Luise, Cristian Lussana, Rolf Heilemann Myhre, Amélie Neuville, Even Marius Nordhagen, Thomas N. Nipen, Sorcha Owens, Steven Ramsdale, Jesica Pinon Rodriguez, Ivar A. Seierstad, Hakam Shams, Marieke Wesselkamp and Markus Zehner.
Revision number	V1.2
Status	Final
Dissemination Level	Public

Funded by the European Union

The WeatherGenerator project (grant agreement No 101187947) is funded by the European Union.
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. Neither the European Union nor the granting authority can be held responsible for them.

1 Executive Summary

The WeatherGenerator project is developing a data-driven Foundation Model of the Earth system that can support weather, climate and environmental applications. To ensure that the model produces scientifically credible and physically realistic results, a comprehensive evaluation framework has been established within Work Package 3 (WP3). This deliverable presents the first integrated assessment of the WeatherGenerator prototypes across atmospheric forecasting, land-surface applications and explainable artificial intelligence (XAI).

Key findings from the evaluation include:

- **Stable medium range forecasting:** Global forecast verification demonstrates that the current WeatherGenerator atmospheric prototype produces stable medium-range forecasts with performance comparable to leading AI weather prediction systems.
- **Targeted weather system evaluation:** Targeted evaluations show encouraging skill in important application areas, including temperature downscaling, tropical cyclone forecasting and extratropical cyclone prediction, with cyclone forecast errors broadly consistent with those found in contemporary numerical weather prediction systems.
- **Land surface prediction capabilities:** The WeGen-Land prototype demonstrates promising capabilities for Earth-observation reconstruction, gap-filling and forecasting of key land-surface variables, including vegetation state and land-surface temperature.
- **Explainable AI assessment:** Initial XAI investigations reveal physically plausible forecast sensitivity patterns, providing confidence that the model is learning meaningful meteorological relationships and establishing a basis for future interpretability studies.

This deliverable provides both a snapshot of current model performance and the foundation for future evaluation activities within the project. While results are encouraging, the assessments also identify priorities for further development, particularly for long-range weather scenario generation, higher-resolution forecasting and advanced explainability methods. The evaluation framework described here will support the continued improvement of the WeatherGenerator throughout the remainder of the project.

Table of Contents

2 Introduction

2.1 Background

Climate change poses major risks to society's wellbeing and prosperity in Europe and worldwide. Earth system models, which represent interactions between the atmosphere, oceans, land, ice, lakes and chemistry, are essential tools for understanding climate change and associated weather extremes.

The WeatherGenerator project will develop a world-leading generative Foundation Model of the Earth system, serving as a new Digital Twin for Destination Earth (DestinE). Based on representation learning, it will provide a versatile framework for modelling Earth system dynamics by integrating diverse observations and simulations at unprecedented scale.

Bringing together Europe's leading research institutes, SMEs and industry partners in Earth system modelling, high-performance computing and machine learning, the project will create an innovative Digital Twin with applications in the energy, food, water and health sectors.

By advancing both weather and climate science and machine learning, WeatherGenerator will establish a new state of the art in Earth system modelling. Its greater efficiency and flexibility compared with current models will enable faster DestinE services, support testing of management options, and provide enhanced interactive capabilities for users such as policymakers, planners, architects and engineers.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverable

This deliverable contains the preliminary WeatherGenerator evaluation results and is tied to WP3, Task 3.8.

This document summarizes some key features of the model and important physical features (extratropical and tropical cyclones) in global weather and climate. The aim is also to establish a robust scientific evaluation framework for the WeatherGenerator, ensuring its reliability and scientific credibility. It includes the development of a dedicated evaluation tool suite for WeGen-Atmo and WeGen-Land, aligned with WP3 activities.

It will assess the added value of multi-modal training by benchmarking the Foundation Model against uni-modal baselines, including zero-shot performance. For WeGen-Land, the objectives are to evaluate (i) the land assimilation engine through gap-filling experiments and downstream validation, and (ii) the land prediction engine via free-running simulations and physical consistency checks. The evaluation suite will also support continuous testing and validation throughout model development.

2.2.2 Work performed in this deliverable

In this deliverable the work as planned in the Description of Action (DoA, WP3 T3.8) was performed.

Scientific evaluation of the WeatherGenerator is critical for its effectiveness and its acceptance by scientists. It has been performed in parallel to the other tasks of WP3 and encompasses the development of evaluation tools for WeGen-Atmo and WeGen-Land. It will build on existing methodologies from Earth system science but adapted to the challenges and peculiarities of a data-driven Earth system model. In particular, the scientific evaluation will include the use of methods from explainable AI adapted to the very large model sizes of the WeatherGenerator. An emphasis will be on demonstrating the benefits that arise when one trains on a diverse set of data.

The WeGen-Land prototype is evaluated in a twofold manner. First, we assess the land assimilation engine through its masked reconstruction capabilities, demonstrating its ability to recover incomplete Earth observation data and reproduce key spatial features such as land-cover patterns and topographic detail. This evaluation highlights the model's inherent cross-modal data fusion capabilities and its potential for handling missing observations. Quantitative reconstruction metrics will be introduced in future iterations. Second, we evaluate the land forecasting engine by producing forward predictions of key land-surface variables and assessing their skill on downstream forecasting tasks. Current evaluations focus on vegetation indices (NDVI) and land surface temperature (LST), comparing forecasts against observations as well as climatology and persistence benchmarks. The test suite will also be used to monitor performance and ensure functionality during the continued development of the WeatherGenerator.

The evaluation activities planned under WP3 Task 3.8 have been implemented across atmospheric, land-surface and explainable AI applications. Global forecast verification showed that the WeatherGenerator produces stable medium-range forecasts with performance comparable to current AI-based forecasting systems. Targeted evaluations demonstrated encouraging skill in downscaling, tropical cyclone prediction and extratropical cyclone forecasting, with cyclone position and intensity errors generally comparable to those reported for physical weather prediction models. Land-surface evaluations showed promising capability in data reconstruction, gap-filling and forecasting of key variables such as NDVI and land-surface temperature. Explainable AI techniques were also successfully applied to investigate forecast sensitivities and support scientific interpretation of model behaviour. Areas requiring further development were identified, particularly for year-long weather scenario generation, where biases in the seasonal cycle and long-lead-time artefacts remain evident. These results provide the foundation for continued evaluation and model improvement throughout the project.

2.2.3 Deviations and counter measures

No deviations have been encountered

2.2.4 WeatherGenerator Project Partners:

ECMWF	EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS
FZJ	FORSCHUNGSZENTRUM JUELICH
MetNor	NORWEGIAN METEOROLOGICAL INSTITUTE
MPG	MAX-PLANCK-GESELLSCHAFT ZUR FORDERUNG DER WISSENSCHAFTEN E.V.
KNMI	ROYAL NETHERLANDS METEOROLOGICAL INSTITUTE
MetFrance	MÉTÉO-FRANCE
SMHI	SWEDISH METEOROLOGICAL AND HYDROLOGICAL INSTITUTE
UKMO	UK METOFFICE

CMCC	CENTRO EURO-MEDITERRANEO SUI CAMBIAMENTI CLIMATICI
eScience	NETHERLANDS ESCIENCE CENTER
Buluttan	BULUTTAN
KAJO	KAJO SERVICES
LT	LATEST THINKING
Statkraft	STATKRAFT
ETHZ	EIDGENOSSISCHE TECHNISCHE HOCHSCHULE
MetSwiss	METEOSWISS

2.2.5 Weather Generator Experiments, Baseline Datasets, and Evaluation Approaches.

The relevant experimental information, baseline datasets and their associated key variables, and the evaluation approaches are given in Table 2-1.

Application Area	Experiment / Use Case	WeatherGenerator Run	Forecast Timescale	Baseline Dataset(s)	Key Variables	Evaluation / Assessment Approach
Numerical weather prediction **Section 3.1**	Medium-range weather forecasting (Oct–Dec 2023)	atmosfs03	10 days	ERA5	Standard atmospheric forecast variables	RMSE and correlation plots using WeGen FastEvaluation and Quaver.
Reanalysis downscaling **Section 3.2.1**	Downscaling reanalysis products	atmosfs02/03	Variable	ERA5, CERRA, MEPS, NORA3	2 m temperature	Comparison against observational and reanalysis reference datasets.
Tropical cyclone forecasting **Section 3.2.2**	Case-study forecasts of tropical cyclones	atmofs02	7 days	ERA5	Sea-level pressure, 10 m wind speed	Assessment of cyclone intensity, structure and evolution relative to reanalysis.
Extratropical cyclone forecasting **Section 3.2.3**	Storm detection, tracking and forecast matching	atmofs03	Up to 10 days	ERA5	Mean sea-level pressure, 850 hPa zonal and meridional winds	Cyclone identification, tracking and matching using TempestExtremes and Huracanpy.
Hydrological applications **Section 3.2.4**	Generation of seasonal-to-multi-year weather scenarios	atmofs03	Seasonal to multi-year	ERA5	Primarily 2m temperature (precipitation planned for future releases)	Long-range verification and evaluation of weather scenarios for hydrological use.
Land-surface forecasting. **Section 3.3**	Gap-filling Earth observation data and forecasting land processes	Land Prototype	7 days (sub-daily variables), 30 days (daily variables)	Sentinel-2, SEVIRI, ERA5	Land surface temperature, skin temperature, NDVI, vegetation activity	Comparison with satellite observations and benchmark models for vegetation and surface state prediction.
Explainable AI (XAI) **Section 3.4**	Model interpretability and sensitivity analysis	atmofs02	Case-study based	ERA5	300 hPa geopotential height (input), MSLP (target variable)	Gradient × Input relevance analysis and visualization of forecast drivers.

Table 2-1: Relevant data and model configurations for the evaluation section.

3 Evaluations

The following section contains all the evaluations performed by the co-authors of this document. An overview of model performance is provided in Section 3.1; targeted assessments are provided in Section 3.2; the assessment of WeGen-Land is provided in Section 3.3; and an overview of progress on XAI is given in Section 3.4.

3.1 The WeGen FastEvaluation (atmosphere) – global scores and maps: Matthias Karlbauer and Ilaria Luise, ECMWF

3.1.1 Introduction and results summary

Global scores and maps of the WeatherGenerator forecast model released in March 2026 are presented in this section. The WeGen FastEvaluation tools are used to provide the plots in this section (see the official documentation at this link).

Key results

The model is performing as expected and is comparable in performance to PanguWeather. The model is stable for long rollouts of at least a month. Furthermore, the model does not diverge for medium range forecasting. Finally, the 2D scores enable the areas of higher errors to be identified.

3.1.2 Overall performance scores

Plots of the root mean squared error (RMSE) and anomaly correlation coefficient (ACC) computed relative to ERA5 are presented in this section to provide a baseline assessment (see Figure 3-1 and Figure 3-2). Performance is compared against the other AI models (GraphCast, PanguWeather and AIFS). Please note that GraphCast, IFS and AIFS are scored against IFS analyses, while PanguWeather and WeatherGenerator are scored against ERA5. Further plots of performance in the Tropics and Southern Hemisphere are provided in the deliverable folder[1]. The model has skill comparable with the existing AI models. All scores are computed on a [1.5, 1.5] grid to comply with WM standards. Note: in Figure 3-1 and Figure 3-2, the ERA5 analyses (type="an") are compared against the ERA5 forecasts (type="fc") retrieved from the MARS catalogue[2].

[1] WG_model_atmosfs03-ACC.pdf
WG_model-atmosfs03-RMSE.pdf

[2] https://apps.ecmwf.int/mars-catalogue/

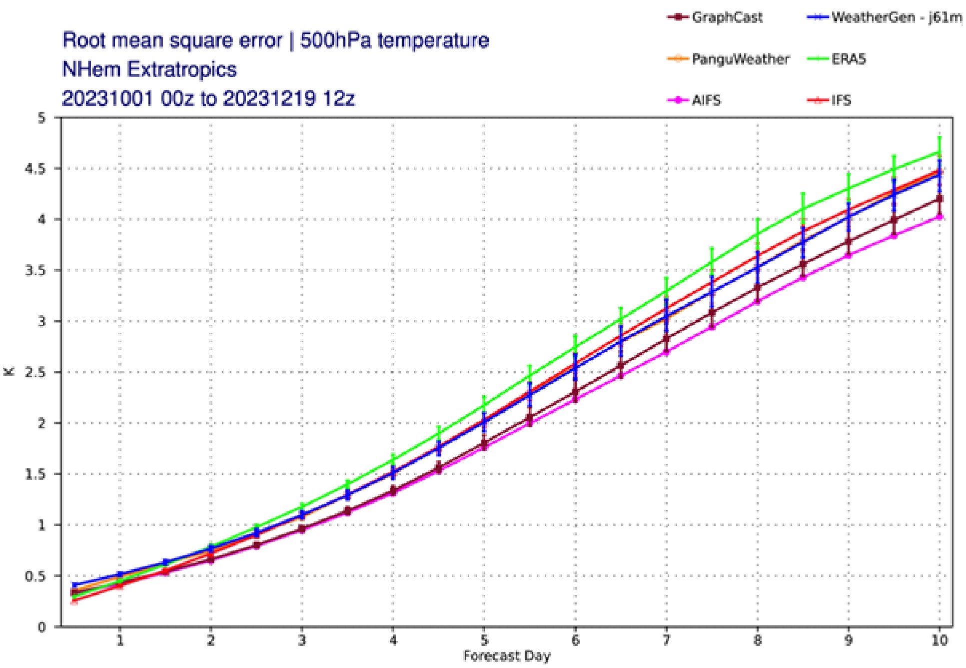

Figure 3-1: 500 hPa air temperature RMSE in the Northern Hemisphere extratropics compared to the other forecasting models GraphCast, PanguWeather, AIFS and IFS.

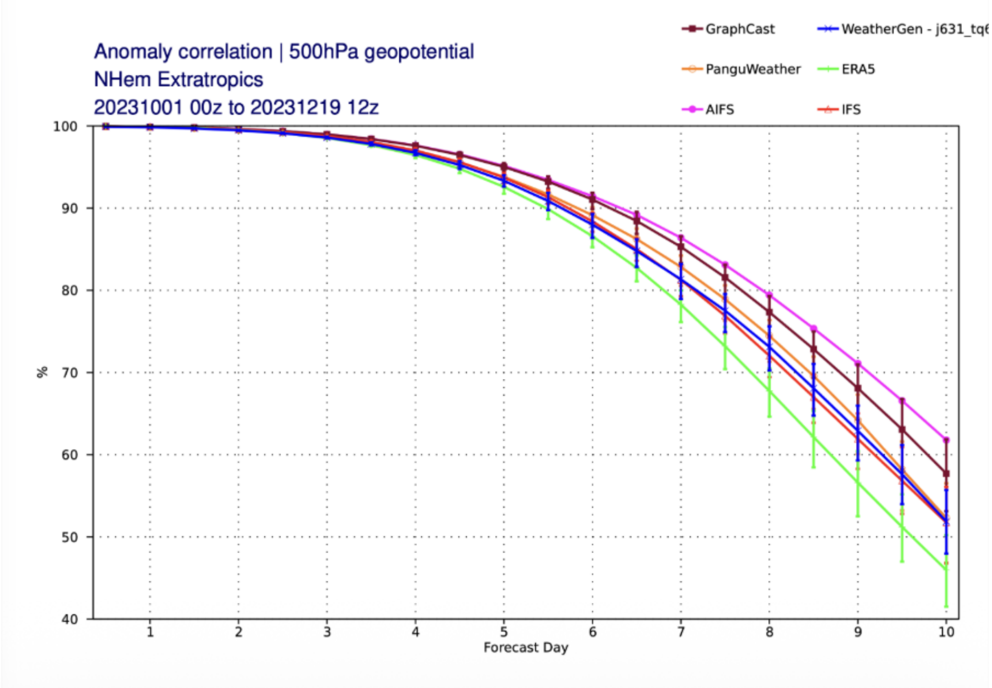

Figure 3-2: 500 hPa geopotential anomaly correlation over the Northern Hemisphere Extratropics compared to the other forecasting models GraphCast, PanguWeather, AIFS and IFS.

3.1.3 2D Maps

Thanks to the WeGen FastEvaluation tool[3] we can inspect the structure of the predicted fields. Examples of prediction vs target plots at 72 hours from forecast initialization can

[3] https://wgen-fasteval.readthedocs.io/en/latest/index.html

be seen in Figure 3-3. Note that 72 hours corresponds to 12 steps of 6 hours. The plots show good performance at 72 hours forecast. Biases are generally localised rather than concentrated in specific regions, and for most variables they remain within a few percent of the total value. This holds for both smooth fields, such as temperature and geopotential, and more variable fields, such as specific humidity (q).

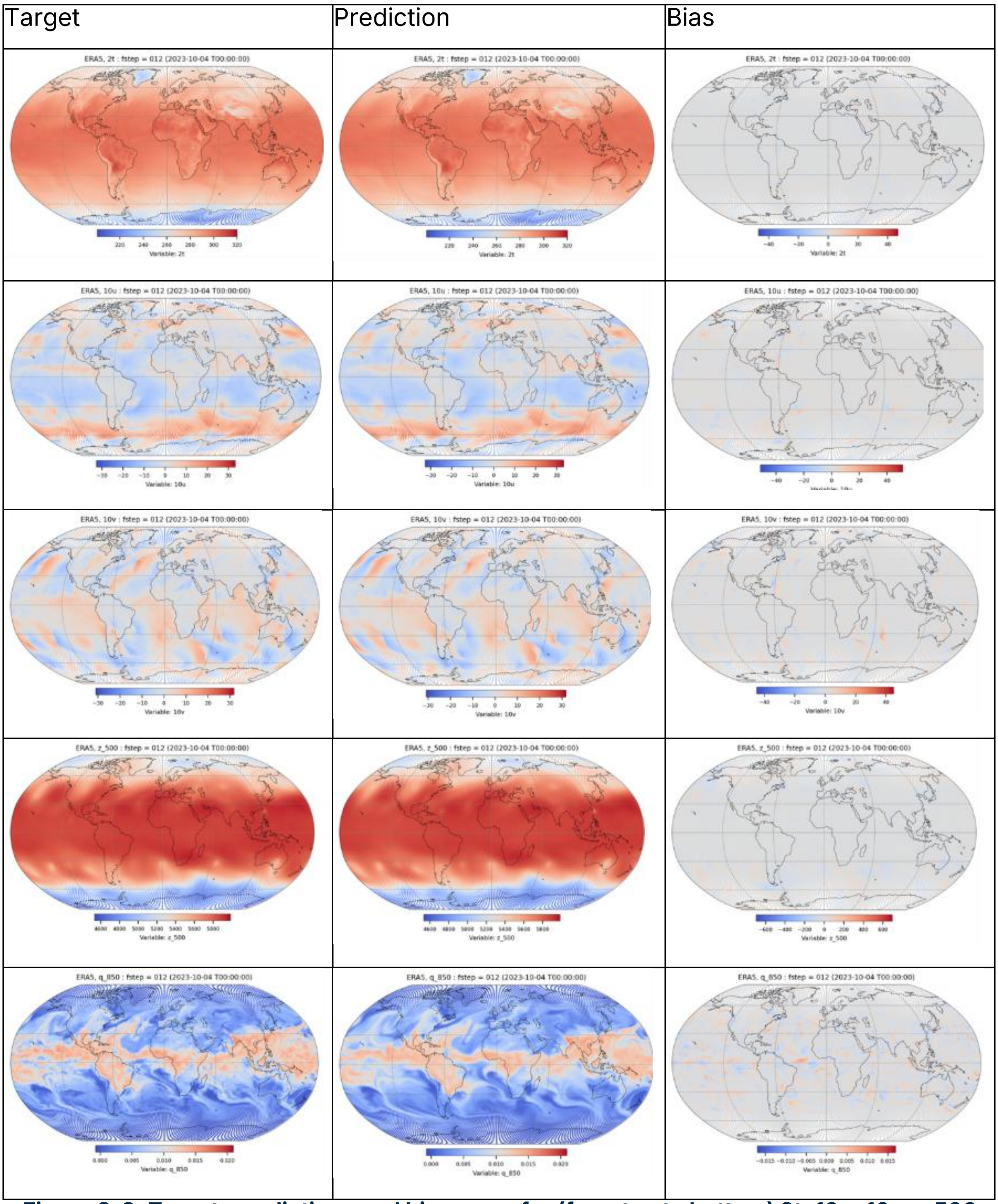

Figure 3-3: Target, predictions and bias maps for (from top to bottom) 2t, 10u, 10v, z 500 and q 850 for a 72h forecast with the WeatherGenerator model. The high resolution version of these plots can be found here: high-resolution plots.

3.1.4 2D Score plots

With the WeGen FastEvaluation package we can also plot the 2D distribution of the RMSE scores for each forecast step. This is useful to see if the degradation in the RMSE scores occurs over specific areas of the globe and therefore indicates a regional problem with the model configuration.

The RMSE plots (Figure 3-4) shows that the error is higher on land than on the ocean. The largest RMSE values are also correlated to the vertical profile of the land surface (e.g. Andes and Himalaya) and the high RMSE values are also co-located with areas of sea-ice coverage at the poles. All plots seem to suggest higher error coverage at the poles.

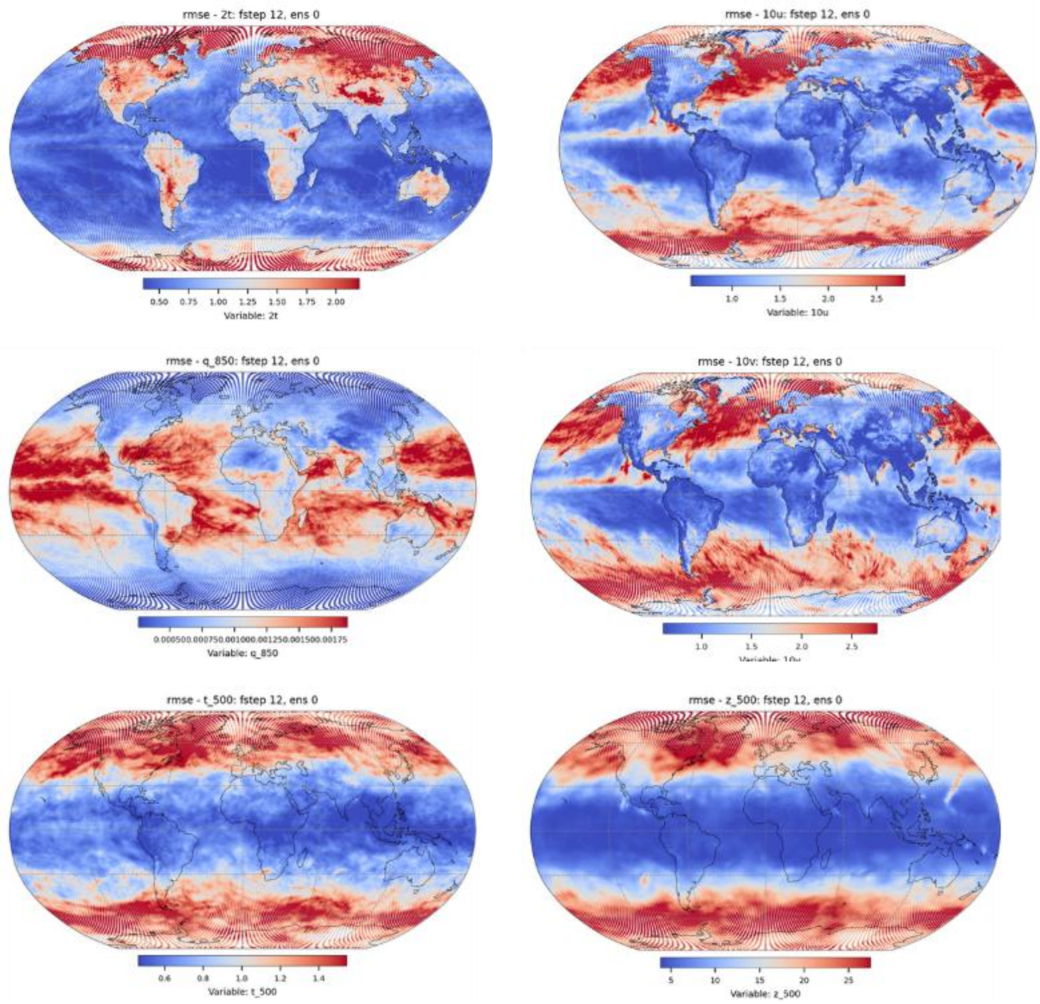

Figure 3-4: RMSE error averaged over 128 initialization times in space for 2t, 10u, q 850, 10v, t 500, z 500 for a 72h forecast with the WeatherGenerator. The red parts imply higher error rates on average than the blue parts. The high resolution version of these plots can be found here: high-resolution plots.

The average error maps for the different fields at 72 hours generally show larger temperature errors near the poles and smaller errors near the equator (Figure 3-4). This could be improved, for example, by refining the latitude weighting. For surface temperature (Figure 3-4, top row), the errors also appear linked to the sea-ice interface, although sea-ice variables are not currently included in the training data. For specific humidity (Figure 3-4, middle row), errors are higher in regions of strong convection, such as the tropics, as expected.

Overall, the plots indicate that errors remain within a few percent of the total value on average, with no concerning spatial patterns. As expected, the errors are concentrated mainly in regions of strong convection or instability. No major concerns arise from this first round of evaluation of the deterministic forecast model.

3.2 Targeted Evaluations (atmosphere)

3.2.1 Near Surface Fields – 2m Temperature over Europe: Cristian Lussana, Rolf Heilemann Myhre, Amélie Neuville, Even Marius Nordhagen, Ivar A. Seierstad and Thomas N. Nipen, Norwegian Meteorological Institute

3.2.1.1 Introduction and results summary

The focus of this section is to evaluate the WeatherGenerator as a statistical downscaling tool, with the goal of downscaling two global and continental reanalysis products: ERA5 (spatial resolution ~100 km) and CERRA (5.5 km); to predict the output of our high-resolution numerical weather prediction model MEPS (2.5 km – see Frogner et al., 2019). The primary objective of this evaluation is to quantify the downscaling skill of the WeatherGenerator in this configuration. A secondary, longer-term objective is to exploit the multi-stream training capabilities of the WeatherGenerator to blend outputs from multiple numerical models and decode them into target datasets of particular interest, such as observational datasets or regional reanalyses over the Nordic domain.

The results show the process is simulated adequately at this stage of the project and while the performance is adequate, further work is needed to improve the output as the project progresses.

3.2.1.2 Detailed overview

The model was trained over a three-year period using a Hierarchical Equal Area isoLatitude Pixelization (HEALPix, see Górski et al., 2005) level 5 grid (12,288 cells covering the globe). The local and global latent space dimensions were set to 1,024 and 2,048, respectively. Downscaling was then applied to the two-meter temperature field for July 2023.

Figure 3-5 presents the July 2023 mean two-meter temperature for three configurations: the MEPS target field (left panel), a baseline consisting of ERA5 bilinearly interpolated onto the MEPS grid (middle panel), and the WeatherGenerator output (right panel). The time-averaged verification statistics for July 2023 show that the baseline yields a bias of 0.30 °C and an RMSE of 1.21 °C, while the WeatherGenerator achieves a slightly lower bias of −0.21 °C and a comparable RMSE of 1.20 °C, indicating a marginal but consistent improvement over the baseline.

We are additionally investigating alternative WeatherGenerator configurations. Two aspects are being varied:

(1) the HEALPix resolution used for latent space computation, comparing level 5 (12,288 cells) and level 6 (49,152 cells); and
(2) the inclusion of static geographical predictors during training, comprising more than 40 fields such as surface elevation, land–sea mask, and urban fraction.

To assess the impact of HEALPix resolution, we downscaled ERA5 and CERRA to the NORA3 grid (Haakenstad et al., 2021) and compared results at both HEALPix levels (Figure 3-6). NORA3 has a spatial spacing of 3 km and can be seen as very similar to MEPS (in terms of the modelling chain used) but it is an hindcast (dynamical downscaling of ERA5). In this experiment, the model was trained over a ten-year period and evaluated for January 2022. Increasing the HEALPix resolution from level 5 to level 6 yielded a clear improvement, with RMSE decreasing from 0.73 °C to 0.60 °C.

Overall, the results are promising and indicate substantial potential for further improvement. One limitation currently under investigation is the spatial discontinuity of the downscaled fields in certain regions, particularly over the ocean, where tile-like artifacts are apparent. These are likely attributable to the discrete nature of the HEALPix tessellation at the resolution levels tested.

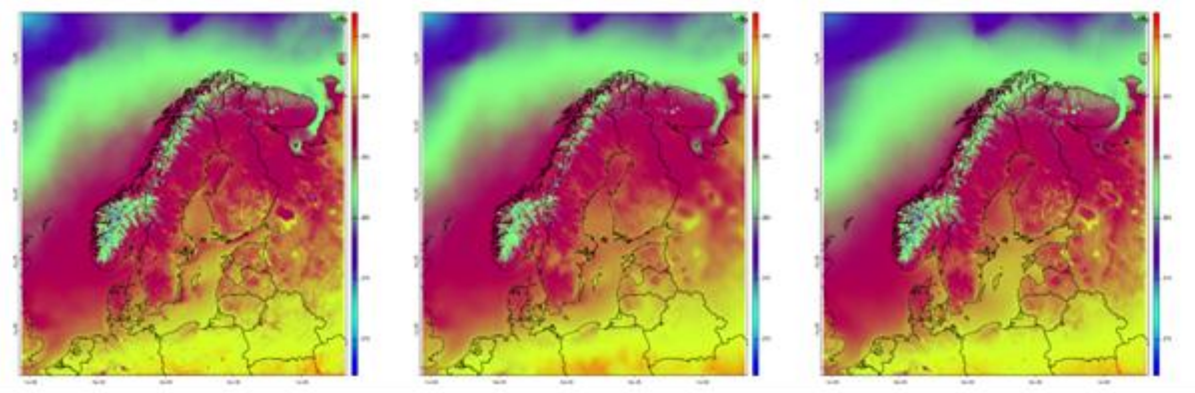

Figure 3-5: Mean July 2023 Two-meter temperature over the MEPS domain. On the left: MEPS (target); in the middle ERA5 baseline; on the right: WeatherGenerator output.

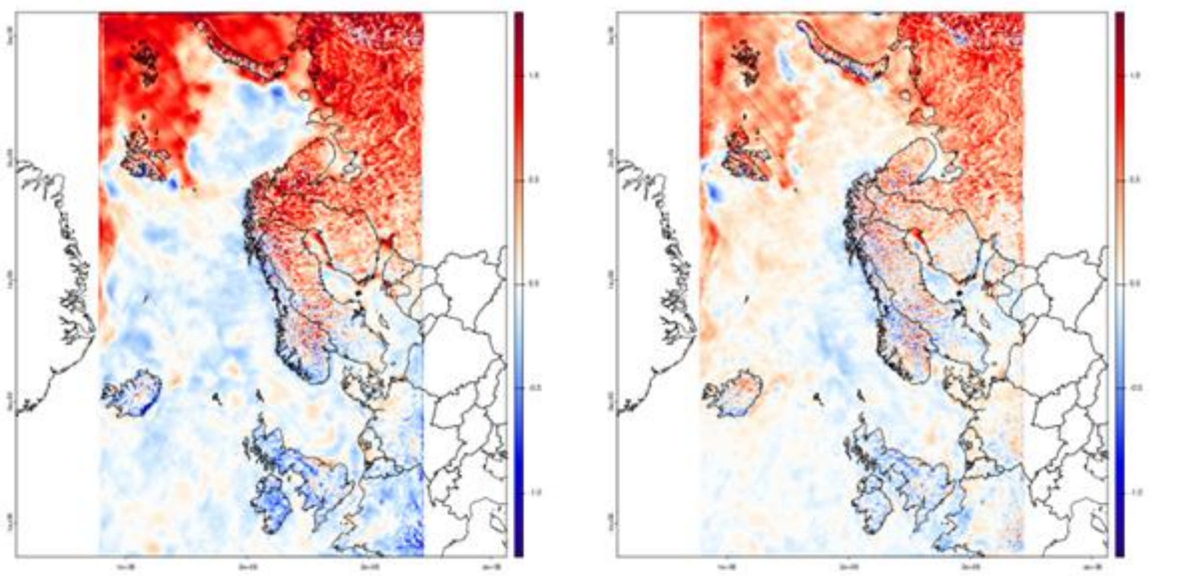

Figure 3-6: Two-meter temperature averaged over January 2022 over the NORA3 domain.

3.2.2 Tropical cyclone case studies: Sebastian Buschow, JSC

3.2.2.1 Introduction and results summary

Tropical cyclones are powerful rotating storm systems that form over warm tropical oceans that are characterised by strong winds and heavy rainfall. They can cause significant damage through flooding, storm surges, and high waves when they reach land. It is therefore very important for these systems to be represented in WeatherGenerator.

From the results in this section, individual tropical cyclones at O96 resolution are visually realistic with good performance in terms of track position, core pressure and wind speeds. A rigorous evaluation using ECMWF tools will be conducted when a higher resolved forecast model (N320, 1h time steps) is ready.

3.2.2.2 Detailed overview

Despite the coarse spatio-temporal resolution (O96 corresponding to approximately 1°x1° in the tropics, 6-hour time steps), a first visual analysis of near surface winds showed coherent looking tropical cyclones in various ocean basins. As an example, Figure 3-7 shows Typhoon Bolaven (from 2023) across a 7-day WeatherGenerator forecast and in the baseline reanalysis data (ERA5).

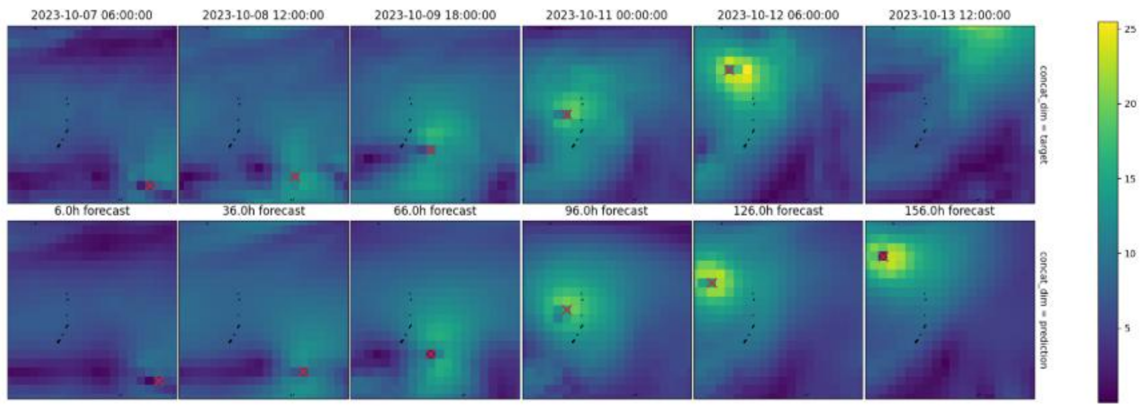

Figure 3-7: ***Snapshots of Typhoon Bolaven in ERA5 (top) and the WeatherGenerator forecast (bottom). Colours correspond to 10m absolute wind speed, the red markers are the storm centre detected by our tool.***

While parts of the simulated wind field are clearly smoothed compared to EAR5 data, the simulated Typhoon itself retains a realistic shape and size, including a single pixel of low wind speeds in the eye of the storm.

For a quantitative assessment of the simulation, the storm was tracked in the forecast and ERA5, using a simple algorithm based on the minima in time of the Laplace-filtered pressure field. The identified storm centers were added to Figure 3-7 as red crosses, confirming that the correct locations were tracked. Figure 3-8 shows the resulting tracks as well as the displacement error, the simulated core pressure and the maximum wind speed in the vicinity of the storm.

We observe a nearly perfect track position for the four days of the simulation, including the transition across the Northern Mariana Islands on October 10, 2023. The intensification in terms of pressure and wind speed is matched adequately as well, with a slight underestimation of maximum wind speed. The extratropical transition after October 13, 2023, is missed by the forecast, resulting in larger track errors near the end of our simulation.

The pressure values and wind speeds are not realistic due to the low spatio-temporal resolution of the target data. It will be interesting to see the performance with respect to extreme wind speeds, once a higher-resolved model version becomes available. At that point, we can start using the ECMWF internal tools and compare physical models and observed best-track estimates like IBTRACS[4].

[4] International Best Track Archive for Climate Stewardship

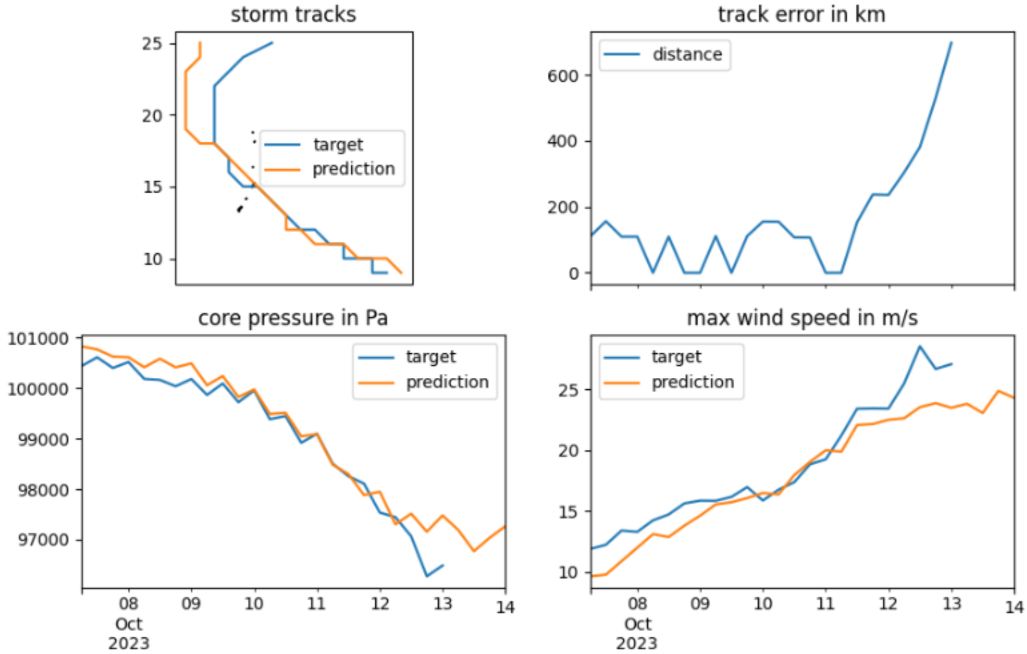

Figure 3-8: ***Strom tracks, track error, core pressure and wind speed of storm Bolaven in the WeatherGenerator forecast and ERA5 target.***

3.2.3 Extratropical cyclones: Duncan Ackerley and Sorcha Owens, UK Met Office

3.2.3.1 Introduction and results summary

Extratropical cyclones are ubiquitous features of the mid-latitude circulation of both hemispheres. They are responsible for many of the hazardous weather features observed at those latitudes and are therefore a key system that needs to be simulated to warn the public (and are routinely named by European Meteorological Services[5]). The analysis in this section matches extratropical cyclones produced by the WeatherGenerator with those that occurred in ERA5. Information on the location, intensity and propagation speed errors can then be calculated as the forecasts evolve. From the dataset given, extratropical cyclone forecasts are within the expected error range of physical weather forecast models (for example see the studies by Froude et al., 2010; 2011) over the same forecast lead times. **Performance is good** and ongoing monitoring is all that is necessary to maintain these results.

3.2.3.2 Detailed overview

The mean errors in separation distance, intensity (central mean sea level pressure) and propagation speed can be seen in Figure 3-9(a), (b) and (c), respectively for Northern Hemisphere extratropical cyclone forecasts by the WeatherGenerator relative to ERA5. There is a monotonic increase in the separation distance error (Figure 3-9(a)) to approximately 300 km by a lead time of 120 hours (5 days). Such an increase is comparable to the biases seen in physical models over the same lead time (see Froude et al., 2010). For central mean sea level pressure (Figure 3-9(b)), there is a tendency for the extratropical cyclones to be too intense in the first 12-36 hours of the forecast before becoming too weak beyond 48 hours into the forecast; however, the mean biases are within ±1 hPa of ERA5 and do not show any persistent over or under deepening during the forecast. Finally, for propagation speed (Figure 3-9(c)), there is a tendency for a

[5] For example, see https://weather.metoffice.gov.uk/warnings-and-advice/uk-storm-centre

negative mean error (cyclones propagate too slowly); however, the 95% confidence interval encompasses the zero line and therefore the bias is unlikely to be significant.

The equivalent mean errors in extratropical cyclones matched between the WeatherGenerator and ERA5 for the Southern Hemisphere can be seen in Figure 3-10. As with the Northern Hemisphere, there is a monotonic increase in the separation distance error for matched Southern Hemisphere cyclones (Figure 3-10(a)) that is consistent with the biases seen in physical models (see Froude et al., 2011). For extratropical cyclone intensity (Figure 3-10(b)), the mean biases are almost zero and therefore Southern Hemisphere cyclones appear to be represented better than their Northern Hemisphere counterparts in the WeatherGenerator. Finally, the mean propagation speed biases are small (relative to ERA5) with the 95% confidence interval overlapping with the zero line. Overall, both the Northern and Southern Hemisphere matched extratropical cyclones are represented in a way that one would expect to see in a physical model.

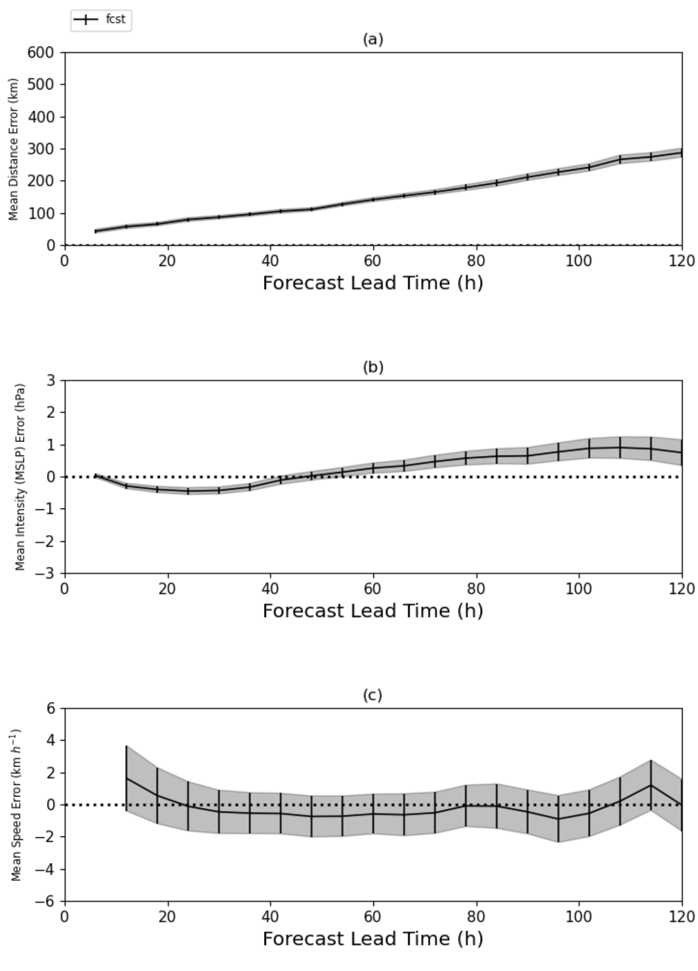

Figure 3-9: Mean Error development of extratropical cyclones in WeatherGenerator vs ERA5 for (a) separation distance (km), (b) intensity difference (mean sea level pressure, hPa) and (c) propagation speed (km hr^{-1}) for the Northern Hemisphere.

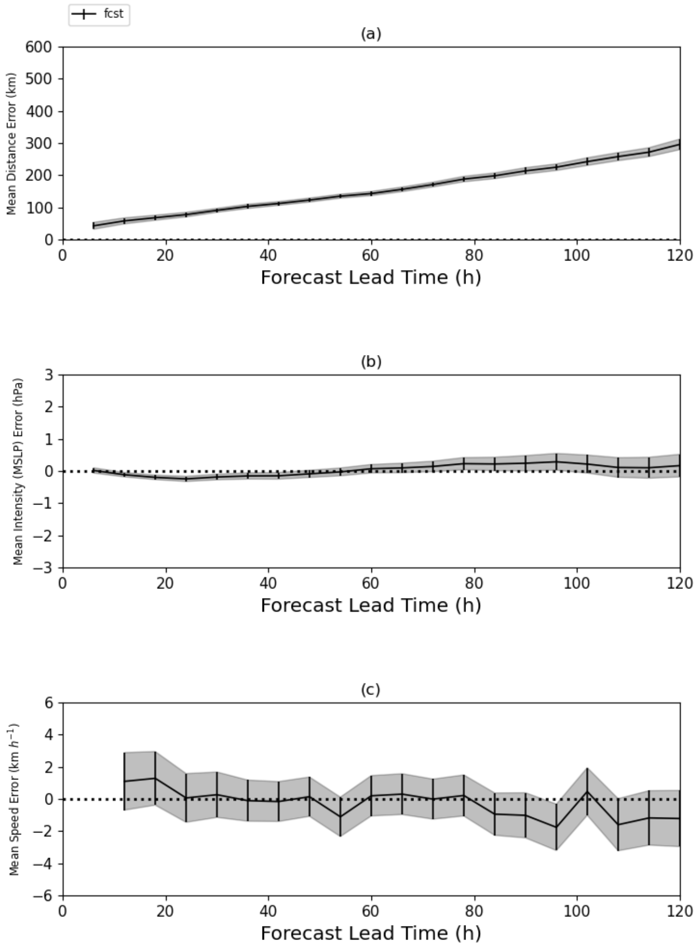

Figure 3-10: Mean Error development of extratropical cyclones in WeatherGenerator vs ERA5 for (a) separation distance (km), (b) intensity difference (mean sea level pressure, hPa) and (c) propagation speed (km hr^{-1}) for the Southern Hemisphere.

3.2.4 Year-long weather scenarios for hydrological applications: Jesica Pinon Rodriguez and Emilie Byermoen, Statkraft; Thomas Nipen and Even Nordhagen, MET Norway.

3.2.4.1 Introduction and results summary

WeatherGenerator was used to generate independent, year-long weather scenarios initialized in January 2023. The models were trained exclusively on ERA5 reanalysis data. Two configurations were evaluated at approximately 100 km horizontal resolution and 6 h temporal resolution, with one generated scenario for each configuration: v1 without dynamic forcings (the release setup), and v2 with dynamic forcings based on geo-information fields (credit: Ankit Patnala, JSC). The generated scenarios were benchmarked against ERA5 averaged over 2003-2022, Statkraft's current model (Martino et al., 2017), represented by 92 scenarios, and the alternative AI model Bris (Nordhagen et al., 2025), represented by 20 scenarios.

Performance is currently not adequate for year-long hydrological scenario generation and needs further focus as the project progresses: the rollouts are stable, but the seasonal temperature cycle is underestimated, and cold artefacts emerge at longer lead times.

3.2.4.2 Detailed overview

Figure 3-11 shows that Bris closely follows the Statkraft scenarios, whereas both WeatherGenerator setups underestimate the seasonal variability at Blåsjø, Norway's largest reservoir by energy storage. In particular, v1 remains too cold and has a weak seasonal cycle, while v2 improves the behaviour but still produces insufficient seasonal amplitude.

Figure 3-12 indicates that the v2 temperature fields remain broadly stable at both 25-day and 250-day lead times, but cold artefacts appear after around 100 days and become more pronounced at longer lead times. This suggests a rollout issue and indicates that long-term dependencies are not yet fully learned. Future work should therefore focus on improved forcing strategies and training approaches for long-term scenario generation.

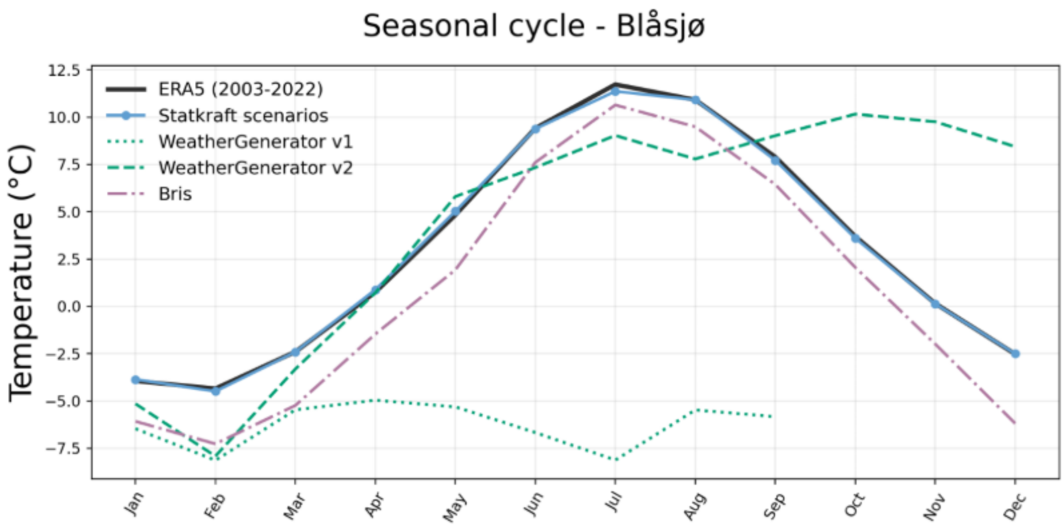

Figure 3-11: Seasonal temperature cycle for Blåsjø, Norway's largest reservoir by energy storage. Bris closely follows the Statkraft scenarios, while both WeatherGenerator setups underestimate seasonal variability.

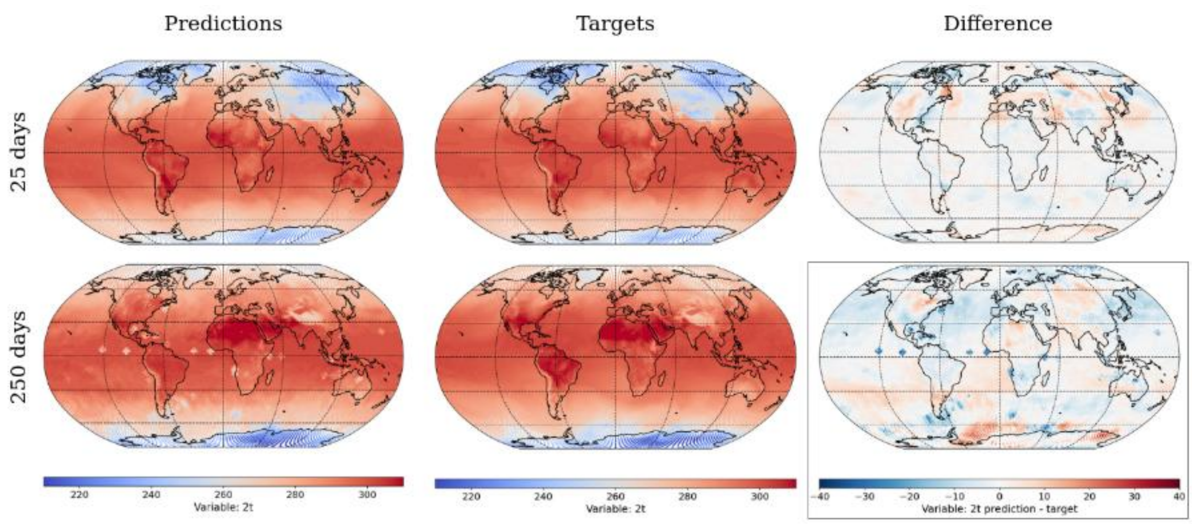

Figure 3-12: WeatherGenerator v2 temperature fields at 25- and 250-day lead times (predictions) and ERA5 fields (targets). Predictions remain broadly stable, but cold artefacts emerge and become more pronounced at longer lead times.

3.3 WeatherGenerator Land: Hakam Shams, Sebastian Hoffmann, Markus Zehner, Vitus Benson, Marieke Wesselkamp, Max-Planck-Institute for Biogeochemistry

3.3.1 Introduction and results summary

In this section, we evaluate JANE (Jena AI model of the Natural Environment), the land component of the WeatherGenerator. Like the atmospheric component JANE is an own data assimilation and forecasting engine, based on masked autoencoding (MAE) (He et al., 2022). Training with this assimilation strategy allows to 1 learn meaningful reconstructions and 2) handle incomplete, i.e. cloudy images. The forecast engine at its current state is trained end-to-end and gets as inputs Sentinel-1 and Sentinel-2. In a transformer with flexible sequence length, it is conditioned on ERA5, Glofas, ERA5 Land during latent space rollout. Subsequent forecasts are decoded into physical space along the conditions over the forecast horizon. Here, we demonstrate JANE for forecasts of two land surface states, that are the normalised difference vegetation index (NDVI), and Land surface temperature (LST). For more details on the model architecture and data pipeline, we refer to D3.3., section 3.

At this stage the model is available as a first prototype. The images that the assimilation engine reconstructs visibly reproduce the spatial context at a level that allows detection of local details, such as topology, land cover types, and shading (Section 3.3.2 below).

However, reconstruction resolution is still low, resulting in an overall blurry impression of the image. This will be improved by more extensive training, normalisation and better masking in the future. Further, a quantification of reconstruction will be conducted. Our end-to-end trained forecast engine already skilfully forecasts land surface temperature (LST) and vegetation indices (NDVI) over medium-range horizons (Sections 3.3.2 below).

3.3.2 Detailed results

3.3.2.1 Gap filling

The reconstruction capability is demonstrated in the example in Figure 3-13 below at the example of a Sentinel-2 RGB image. Panel A) shows the mask that is applied after tokenisation of input images through modality specific encoders. Panel B) shows the image that is reconstructed from a transformer backbone with cross-attention and a Sentinel-2 decoder. Panel C) shows the according ground truth.

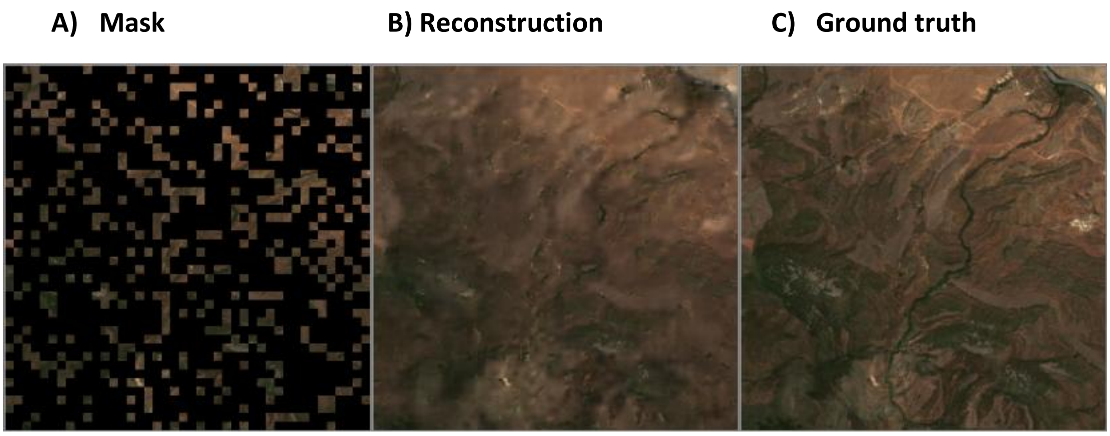

Figure 3-13: ***Example Sentinel-2 RGB image reconstruction (B) from a masked input image (A), compared to the ground truth (C).***

This reconstruction learned MAE can be used for gap-filling, which is especially useful for cloud inpainting. This task is demonstrated in Figure 3-14 with nearest neighbour interpolation, using the Sentinel cloud mask (SCL), compared to OmniCloudMask (Wright et al., 2025). JANE's inpainting skill remains to be demonstrated in the future, as well as a quantified evaluation of the reconstruction task.

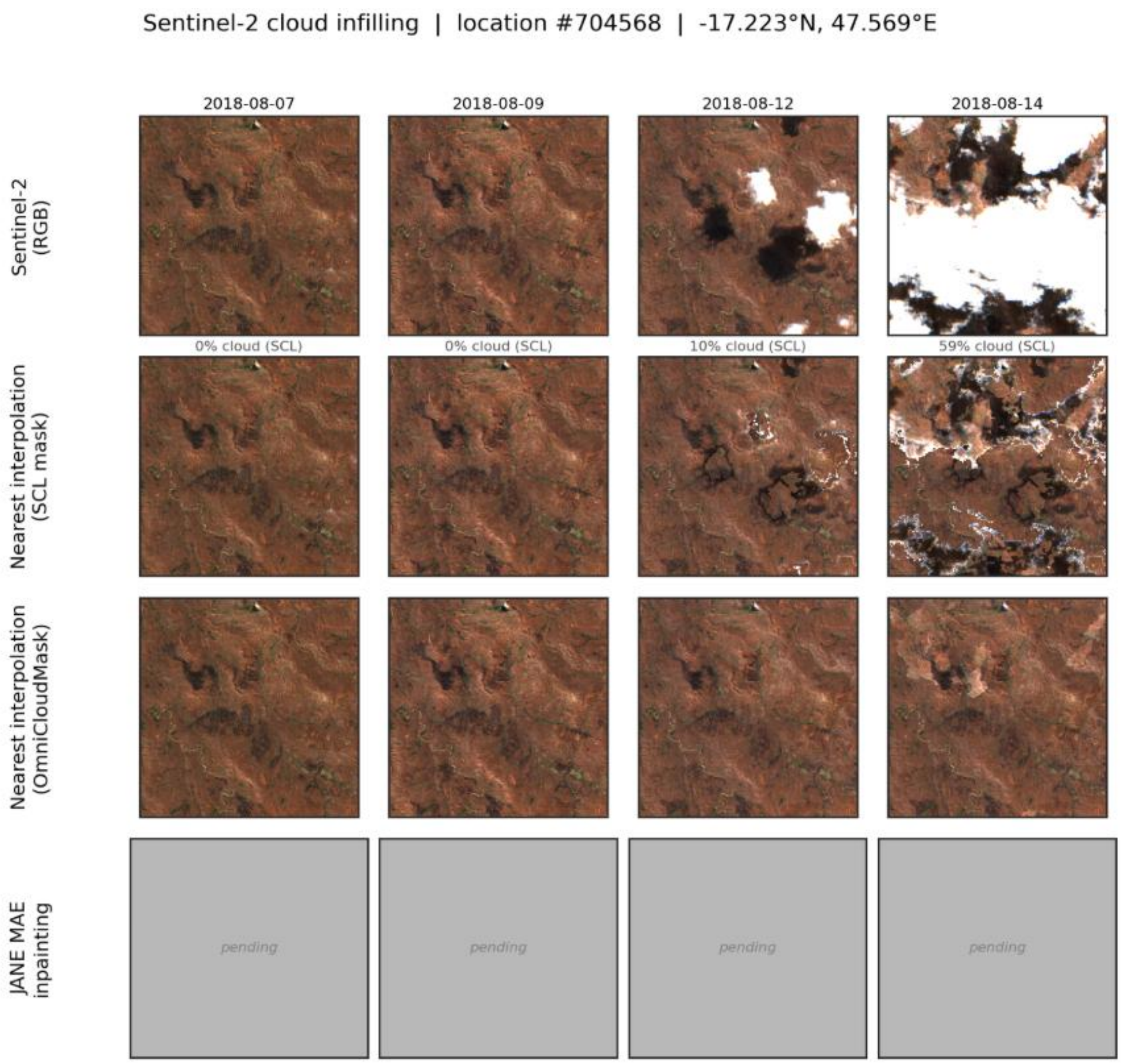

Figure 3-14: ***Cloud inpainting of Sentinel-2 RGB images (top row) with the SCL cloud mask (second row) and OmniCloudMask (third row). This will be compared to JANE in the future (bottom row, pending).***

3.3.2.2 Downstream task 1 – NDVI forecasting

In this section, we demonstrate downstream performance of the forecasting engine. It predicts satellite Sentinel-2 NIR and visible channels on daily resolution (see Figure 3-15) and derives NDVI from it. The model is evaluated over a spatiotemporal test set with 111 pixels in total, at hourly timesteps over the full year 2024. On the example sample in Figure 3-15 it achieves a total RMSE of ~0.08 and MAE of ~0.06. Scores are computed on the available images over the forecast horizon. Total RMSE growth over the forecast horizon, aggregated over all evaluation sites, is shown in Figure 3-16.

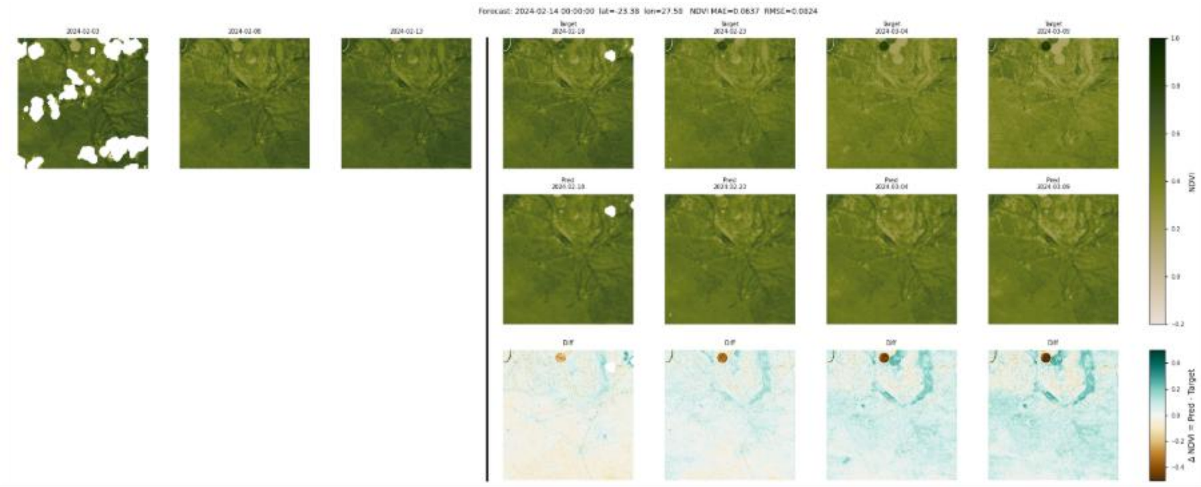

Figure 3-15: ***JANE forecasts of the Sentinel-2 NDVI, derived from its NIR and visible channels (top and middle row). Four images are available at irregular intervals over the forecast horizon of 30 days. The bottom row displays the bias in NDVI predictions.***

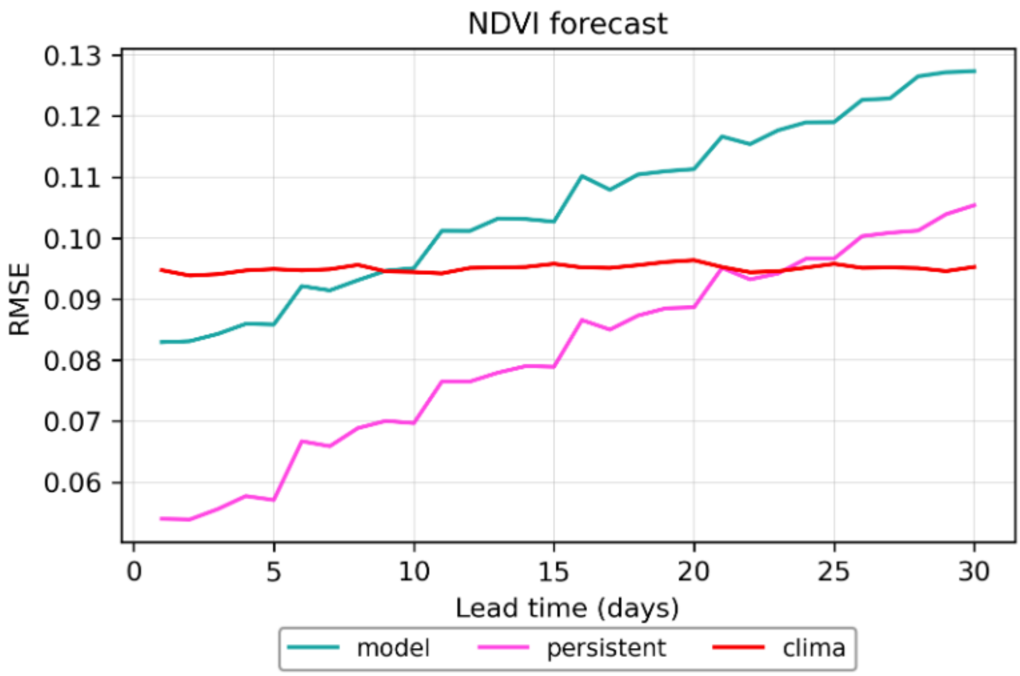

Figure 3-16: ***JANE RMSE growth over a forecast horizon of 30 days, compared to persistence and climatology.***

3.3.2.3 Downstream task 2 – LST forecasting

The decoder for LST is conditioned on the full ERA5 and ERA5-Land datasets, including skin temperature. This means, right now it acts like a bias correction from skin temperature to LST, rather than a pure forecast. The model is evaluated over the same spatiotemporal test set pixels for LST as for NDVI. Over a medium range horizon, hourly forecasts are consistently skilful when compared to the hourly smooth climatology and persistence (see Figure 3-17). This holds true during both, peak day- and night-times. The day-to-day anomalies are well predicted during the day and night (see ACC in all hour-month cells > 0.5), despite large absolute daytime errors especially from October to December and from February to March (see Figure 3-18).

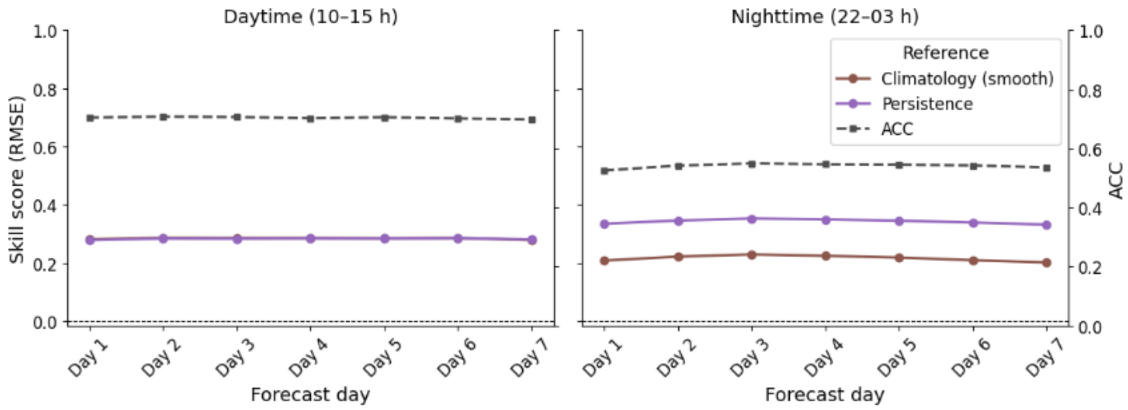

Figure 3-17: ***RMSE skill score of JANE LST forecasts compared to 2 baselines (hourly persistence and smooth climatology). ACC is the anomaly correlation coefficient, computed with the smooth climatology. Left: Skill or validation daytime hours (UTC), Right: Skill for validation nighttime hours (UTC).***

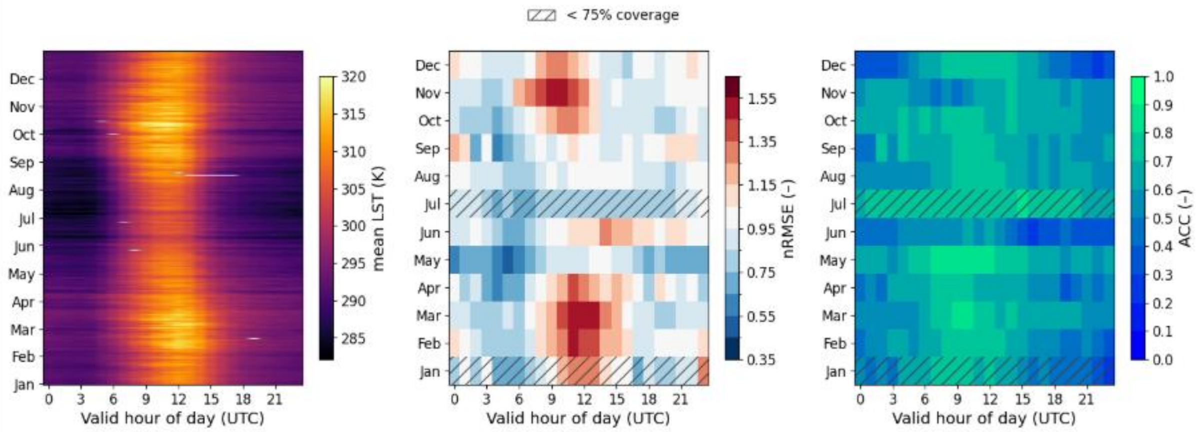

Figure 3-18: ***Left: Fingerprint of observed SEVIRI ST over validation sites and full test year 2024. Middle: RMSE normalised by the global, inter-daily LST variability, pooled over pixels, initialisation and lead times. Red indicates higher error than the usual day-to-day variability in LST, blue lower. Right: ACC, grouped by month and pooled over lead times. Higher ACC is better anomaly prediction, here all cells show an ACC > 0.5. The hatch in both panels indicates months with missing validation data % > 75.***

3.4 XAI: Anna-Louise Ellis, Sorcha Owens and Steven Ramsdale UK Met Office.

To better understand how machine-learning weather models make their predictions, we applied explainable AI (XAI) techniques to forecasts of Storm Goretti. The aim was to identify which atmospheric features were most influential in predicting mean sea-level pressure (MSLP) at Stepper Point, Cornwall, during the storm.

For the WeatherGenerator model, we used a technique called gradient × input sensitivity analysis, which highlights areas of the atmosphere that the model is particularly sensitive to when making a forecast. This work is a first step towards applying a more advanced method called Layer-wise Relevance Propagation (LRP), which can provide a more complete explanation of model behaviour.

To help assess whether the WeatherGenerator sensitivity patterns are meteorologically realistic, we compared them with results from the ECMWF AIFS-Single model. For AIFS-Single, the full Conservative Propagation Layer-wise Relevance Propagation (CP-LRP) method was successfully applied using Microsoft Azure ML. These results provide a useful reference because they offer a more complete picture of how atmospheric features contribute to the forecast.

The WeatherGenerator analysis produced physically plausible and focused sensitivity patterns for Storm Goretti on the HEALPix grid. However, the work remains preliminary because running the full CP-LRP method on the WeatherGenerator currently exceeds the available GPU memory on the JUPITER system. Resolving this technical limitation is the next step towards a more complete interpretability assessment.

In contrast, both gradient × input sensitivity analysis and full CP-LRP were successfully applied to AIFS-Single on a Microsoft Azure ML H100 GPU cluster. The larger memory available on these GPUs is the most likely reason why the full LRP calculations could be completed for AIFS-Single but not yet for the WeatherGenerator. Identifying whether optimisations could be implemented to reduce the memory consumed by the WeatherGenerator model is also being explored and would benefit both the TRUST_ML repository and users of the WeatherGenerator model. As a result, the AIFS-Single relevance maps are used as a physically based benchmark against which the WeatherGenerator results can be compared.

3.4.1 Detailed overview

Figure 3-21(a) shows the observed atmospheric conditions during Storm Goretti at 18:00 UTC on 8 January 2026 using ERA5 reanalysis data. These observations provide the meteorological context needed to interpret the explainable AI results.

Figure 3-19 shows the WeatherGenerator gradient × input sensitivity analysis for a 6-hour forecast of mean sea-level pressure at Stepper Point. The analysis examines how variations in upper-atmosphere height at 300 hPa (Z300) influence the forecast pressure. Areas shown in warm colours indicate locations where higher Z300 values contribute positively to the pressure forecast, while cool colours indicate locations with a negative influence. It is important to note that this method shows where the model is sensitive to changes in the input data, rather than providing a direct measure of cause and effect. The dotted appearance of the map reflects the WeatherGenerator architecture, which only processes a subset of grid points during each forecast.

Figure 3-20 shows the equivalent analysis for AIFS-Single using the full CP-LRP method. Unlike gradient × input sensitivity, CP-LRP tracks how information flows through the entire model and provides a more complete picture of which atmospheric features contribute to the forecast. Warm colours indicate positive contributions to the predicted pressure at Stepper Point, while cool colours indicate negative contributions.

The ERA5 analysis shows that the surface low-pressure system had crossed the upper-level jet stream by this time. Such a configuration is often associated with rapid intensification of extratropical cyclones. The extensive region of positive relevance shown in the AIFS-Single results (Figure 3-20) is consistent with this meteorological process and suggests that the model is linking the jet-stream interaction to the pressure changes observed at Stepper Point six hours later.

The negative relevance region over the central North Atlantic in Figure 3-20 appears to be associated with a developing ridge ahead of another upstream cyclone. This ridge could influence the shape and position of the jet stream downstream, affecting the subsequent development of Storm Goretti. Overall, many of the positive and negative relevance patterns highlighted by AIFS-Single correspond to known large-scale atmospheric features that are important for cyclone development and impacts over south-west England.

Figure 3-21(b) shows ERA5 300 hPa wind speeds and mean sea-level pressure at 18:00 UTC on 8 January 2026 and helps to show how the higher-level winds interact with the surface low. When comparing the relevance patterns with the 300 hPa wind speed fields in Figure 3-21(b), the positive relevance region over south-west England in the AIFS-Single results does not coincide exactly with the strongest jet-stream winds. This signal may instead be related to a dry intrusion wrapping around the cyclone centre. By contrast, the strongest WeatherGenerator sensitivity values are more closely aligned with the core of the upper-level jet stream.

Despite the differences between the two explainability methods, comparing Figure 3-19 and Figure 3-20 with the observed meteorological situation shown in Figure 3-21 provides an important preliminary assessment of the WeatherGenerator results. The sensitivity patterns produced by the WeatherGenerator appear to capture physically meaningful atmospheric structures associated with Storm Goretti. Applying full CP-LRP to the WeatherGenerator, once the current GPU memory limitations are resolved, will provide a more definitive evaluation of these initial findings.

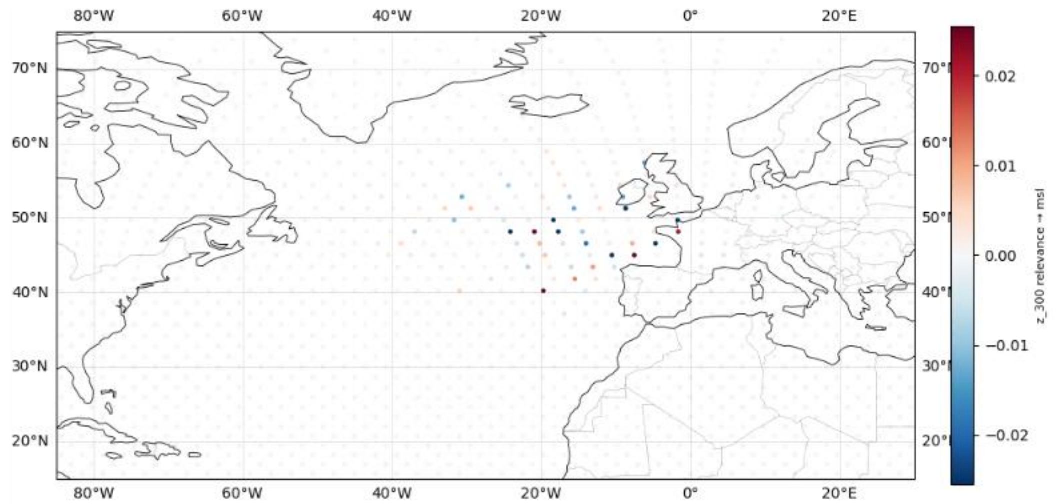

Figure 3-19: ***Gradient × input sensitivity ($rel = x \times \partial y_p/\partial x$) of WeatherGenerator (atmofs02) 6-hour forecast of MSL at Stepper Point, Cornwall (50.05°N, 5.16°W) with respect to input z_300; initialised 2026-01-08T12:00 UTC, valid 2026-01-08T18:00 UTC. Active HEALPix NSIDE=32 (NESTED) nodes, North Atlantic domain, token-size dimension summed. Warm/cool: positive/negative gradient-weighted z_300 activation. Not a conserved decomposition; gaps reflect the model's native grid.***

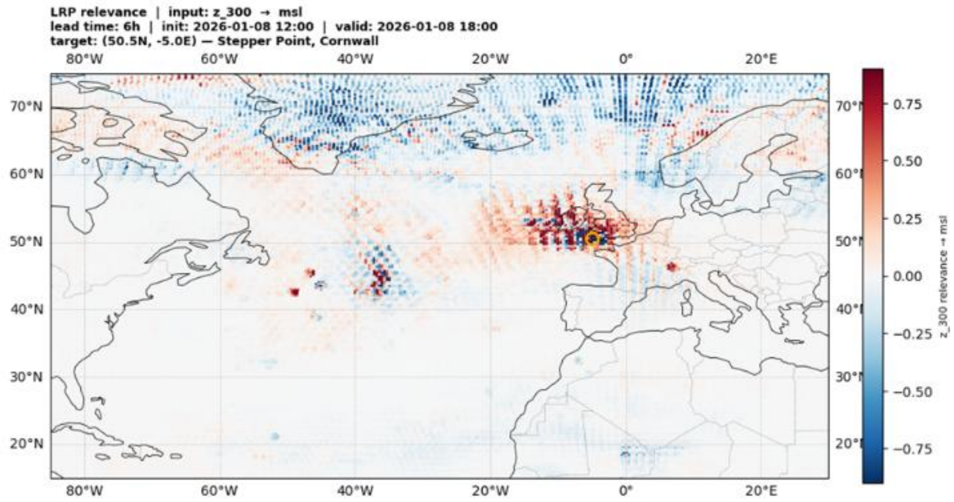

Figure 3-20: ***CP-LRP relevance (R(g) = z_300(g) × δ_LRP(g);Achtibat et al., 2024) of AIFS-Single (ECMWF/AIFS-single-1.0) 6-hour forecast of MSL at Stepper Point (50.05°N, 5.16°W) with respect to input z_300; initialised 2026-01-08T12:00 UTC, valid 2026-01-08T18:00 UTC. Full N320 reduced Gaussian grid (~0.25°), North Atlantic domain [15–75°N, 85°W–30°E]. δ_LRP is the LRP-redistributed relevance coefficient (not a standard gradient); conservation enforced at every layer. Warm/cool: positive/negative conserved contribution to MSL prediction.***

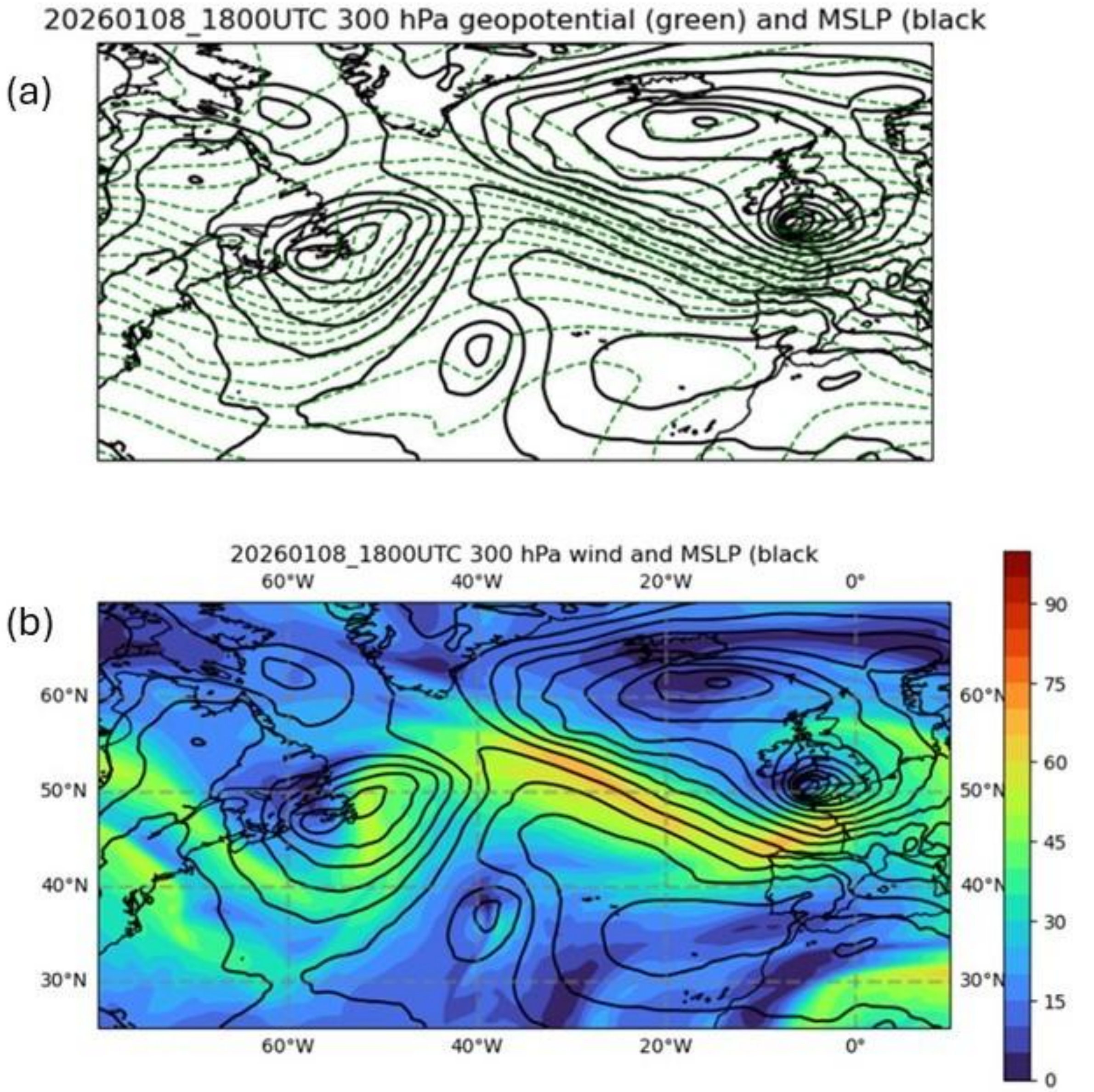

Figure 3-21: ***ERA5 reanalysis at 2026-01-08T18:00 UTC for Storm Goretti. (a) Black contours: MSLP. Green contours: Z_300. (b) Black contours: MSLP. Colours: 300 hPa wind speed. The ERA5 reanalysis gives the best estimate of the observed atmospheric state providing synoptic context for the ML.***

4 Conclusion

This deliverable presents the first comprehensive scientific evaluation of the WeatherGenerator across atmospheric forecasting, land-surface applications and explainable AI diagnostics. Overall, the results demonstrate that the prototype can produce stable forecasts and physically realistic weather evolution across a range of spatial and temporal scales.

For atmospheric forecasting, global verification showed performance comparable to other leading AI forecasting systems, with stable forecast rollouts extending to at least one month. In the extratropical cyclone evaluation, matched storms exhibited position errors of approximately 300 km after five days, consistent with those reported for operational numerical weather prediction systems, while intensity biases generally remained within ±1 hPa. Tropical cyclone case studies similarly demonstrated realistic storm structure, track evolution and intensity changes over forecast lead times of up to seven days.

Downscaling experiments for near-surface temperature over Northern Europe showed measurable improvements relative to interpolated reanalysis baselines, with RMSE reduced from 0.73 °C to 0.60 °C when higher-resolution HEALPix representations were used. For land applications, the WeGen-Land prototype successfully reconstructed missing Earth observation data and produced skilful forecasts of vegetation indices and land-surface temperature, demonstrating the potential of multimodal learning for environmental prediction tasks.

The evaluation also identified important areas for future development. Long-range weather scenario generation remains challenging, with the current prototype underestimating seasonal temperature variability and developing cold artefacts after approximately 100 days. Furthermore, there are HEALPix artifacts at high resolution that are visible when predicting MEPS (see Section 3.2.1). Explainable AI analyses nevertheless showed physically plausible sensitivity patterns linked to known atmospheric processes, providing confidence that the model is learning meaningful meteorological relationships.

Taken together, these results indicate that the WeatherGenerator has already achieved encouraging levels of physical realism and forecast skill at this stage of development, while also providing clear guidance for future model improvements and evaluation activities.

5 References

Achtibat, R., Hatefi, S., Dreyer, M., Jain, P., Wiegand, T., Samek, W., and Lapuschkin, S. (2024). AttnLRP: Attention-aware backpropagation for transformer interpretability. *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*.

Frogner I-L, Singleton AT, Køltzow MØ, Andrae U. Convection-permitting ensembles: Challenges related to their design and use. *Q J R Meteorol Soc*. 2019;145 (Suppl. 1):90–106. https://doi.org/10.1002/qj.3525

Froude, L.S.R. (2010). TIGGE: Comparison of the prediction of the Northern Hemisphere extratropical cyclones by different ensemble prediction systems. Weather and Forecasting, 25, 819-836.

Froude, L.S.R. (2011). TIGGE: Comparison of the prediction of the Southern Hemisphere extratropical cyclones by different ensemble prediction systems. Weather and Forecasting, 26, 388-398.

Górski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M. and Bartelmann, M., 2005: HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *The Astrophysical Journal*, **622**, 2, 759, https://doi.org/10.1086/427976.

Haakenstad, H., Ø. Breivik, B. R. Furevik, M. Reistad, P. Bohlinger, and O. J. Aarnes, 2021: NORA3: A Nonhydrostatic High-Resolution Hindcast of the North Sea, the Norwegian Sea, and the Barents Sea. *J. Appl. Meteor. Climatol.*, **60**, 1443–1464, https://doi.org/10.1175/JAMC-D-21-0029.1.

He, Kaiming, et al., (2022). "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Lang, S., et al., (2024). AIFS — ECMWF's data-driven forecasting system. https://doi.org/10.48550/arXiv.2406.01465.

Martino, S., Nipen, T.N., Lussana, C., and Kolberg, S. (2017). SINTEF Energi AS, ISSN: 1504-9795

Nordhagen, E. et al., (2025). High-Resolution Probabilistic Data-Driven Weather Modeling with a Stretched-Grid, arXiv:2511.23043.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. *ICLR 2014 Workshop*.

Ullrich, P. A., Zarzycki, C. M., McClenny, E. E., Pinheiro, M. C., Stansfield, A. M., and Reed, K. A.: TempestExtremes v2.1: a community framework for feature detection, tracking, and analysis in large datasets, Geosci. Model Dev., 14, 5023–5048, https://doi.org/10.5194/gmd-14-5023-2021, 2021.

Wright, N., Duncan, J.M., Callow, J.N., Thompson, S.E., and George, R.J. (2025). Training sensor-agnostic deep learning models for remote sensing: Achieving state-of-the-art cloud and cloud shadow identification with OmniCloudMask. *Remote Sensing of Environment*, *322*, 114694.

Document History

Version	Author(s)	Date	Changes
1.0	Duncan Ackerley, Vitus Benson, Sebastian Buschow, Emilie Byermoen, Anna-Louise Ellis, Sebastian Hoffmann, Matthias Karlbauer, Ilaria Luise, Cristian Lussana, Rolf Heilemann Myhre, Amélie Neuville, Even Marius Nordhagen, Thomas N. Nipen, Sorcha Owens, Steven Ramsdale, Jesica Pinon Rodriguez, Ivar A. Seierstad, Hakam Shams, Marieke Wesselkamp and Markus Zehner.	7th July 2026	Initial version for review by co-authors
1.1	Duncan Ackerley, Vitus Benson, Sebastian Buschow, Emilie Byermoen, Anna-Louise Ellis, Sebastian Hoffmann, Matthias Karlbauer, Ilaria Luise, Cristian Lussana, Rolf Heilemann Myhre, Amélie Neuville, Even Marius Nordhagen, Thomas N. Nipen, Sorcha Owens, Steven Ramsdale, Jesica Pinon Rodriguez, Ivar A. Seierstad, Hakam Shams, Marieke Wesselkamp and Markus Zehner.	14th July 2026	Version for submission after author edits.

1.2	Duncan Ackerley, Vitus Benson, Sebastian Buschow, Emilie Byermoen, Anna-Louise Ellis, Sebastian Hoffmann, Matthias Karlbauer, Ilaria Luise, Cristian Lussana, Rolf Heilemann Myhre, Amélie Neuville, Even Marius Nordhagen, Thomas N. Nipen, Sorcha Owens, Steven Ramsdale, Jesica Pinon Rodriguez, Ivar A. Seierstad, Hakam Shams, Marieke Wesselkamp and Markus Zehner.	30th July 2026	Updated, final version after review.

Internal Review History

Internal Reviewers	**Date**	**Comments**
Xavier Lapillonne	24/ 07 /2026	Minor comments
Tanya Warnaars	30/ 07/ 2026	Editing and formatting