BAMS In Box

Accelerating Community-Wide Evaluation of AI Models for Global Weather Prediction by Facilitating Access to Model Output

Jacob T. Radford,[a,b] Imme Ebert-Uphoff,[a,c] Jebb Q. Stewart,[b] Kate D. Musgrave,[a] Robert DeMaria,[a] Natalie Tourville,[a] and Kyle Hilburn[a]

KEYWORDS:
Numerical weather prediction/ forecasting;
Model evaluation/ performance;
Artificial intelligence;
Model interpretation and visualization

ABSTRACT: Numerous artificial intelligence-based weather prediction (AIWP) models have emerged over the past 2 years, mostly in the private sector. There is an urgent need to evaluate these models from a meteorological perspective, but access to the output of these models is limited. We detail two new resources to facilitate access to AIWP model output data in the hope of accelerating the investigation of AIWP models by the meteorological community. First, a 3-yr (and growing) reforecast archive beginning in October 2020 containing twice daily 10-day forecasts for *FourCastNet v2-small*, *Pangu-Weather*, and *GraphCast Operational* is now available via an Amazon Simple Storage Service (S3) bucket through NOAA's Open Data Dissemination (NODD) program (https://noaa-oar-mlwp-data.s3.amazonaws.com/index.html). This reforecast archive was initialized with both the NOAA's Global Forecast System (GFS) and ECMWF's Integrated Forecasting System (IFS) initial conditions in the hope that users can begin to perform the feature-based verification of impactful meteorological phenomena. Second, real-time output for these three models is visualized on our web page (https://aiweather.cira.colostate.edu) along with output from the GFS and the IFS. This allows users to easily compare output between each AIWP model and traditional, physics-based models with the goal of familiarizing users with the characteristics of AIWP models and determine whether the output aligns with expectations, is physically consistent and reasonable, and/or is trustworthy. We view these two efforts as a first step toward evaluating whether these new AIWP tools have a place in forecast operations.

DOI: 10.1175/BAMS-D-24-0057.1
Corresponding author: Jacob Radford, jacob.radford@noaa.gov
Manuscript received 22 February 2024, in final form 18 November 2024, accepted 6 December 2024

AFFILIATIONS: [a] Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado; [b] NOAA/Global Systems Laboratory, Boulder, Colorado; [c] Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado

1. Introduction

Over the past 2 years, conditions have aligned for an explosion of artificial intelligence–based weather prediction (AIWP) models for global, medium-range weather forecasting. Namely, the release of the fifth major global reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) (ERA5; Hersbach et al. 2020) in 2020 provided historical weather data at a sufficient resolution for model training, generative artificial intelligence (AI) techniques have advanced and become commonplace, and compute power has progressed to handle extremely large datasets. This has culminated in the release of AIWP models such as FourCastNet (Pathak et al. 2022) and FourCastNet v2-small (Bonev et al. 2023), Pangu-Weather (Bi et al. 2023), and GraphCast (Lam et al. 2023). These AIWP models use no physical equations or parameterizations to make global, medium-range (<10 day) weather forecasts, but rely on access to extremely large datasets such as ERA5 for training the AI model.

Initial verification results are extremely promising, with root-mean-square errors (RMSEs) comparable to or lower and anomaly correlation coefficients (ACCs) comparable to or higher than the ECMWF's High Resolution (HRES) Integrated Forecasting System (IFS). In fact, GraphCast showed superior performance to the HRES for nearly every evaluated parameter, including pressure, temperature, relative humidity, and wind speed/direction, spanning across pressure levels, in terms of RMSE (Lam et al. 2023). While AIWP model training is expensive, AIWP model inference accomplishes forecasting improvements while running orders of magnitude faster than traditional numerical weather prediction (NWP) models. Rolling model forecast skill scores for many AIWP models initialized with either the ERA5 or IFS can be found on the WeatherBench 2 (Rasp et al. 2024) and ECMWF web charts (ECMWF 2023). WeatherBench 2 also offers hindcasts for various AIWP models and periods to provide a starting point for further verification efforts.

The demonstrated forecast skill by AIWP models on its own makes a strong case for government agencies, such as the National Oceanic and Atmospheric Administration (NOAA) and the ECMWF, to consider and plan for the eventual integration of AIWP into their operational modeling processes. Indeed, both agencies show great interest in this technology. The ECMWF has developed their own AI-based weather forecasting model, named AIFS (Lang et al. 2024), and NOAA recently hosted a workshop on AI for NWP (AI4NWP) [NOAA' Physical Sciences Laboratory (NOAA-PSL) 2023] that focused not only on exploring opportunities of AI-based models for NWP but also on the challenges of integrating AIWP into the research-to-operation cycle.

In addition to performance considerations, the dramatically greater efficiency of AIWP compared to traditional NWP could pave the way for a wide range of new forecasting applications, such as extremely large ensemble prediction systems for better-calibrated probability distributions, rapid initial condition sensitivity analysis, and personal, on-demand weather modeling, and certainly others that we have not yet considered.

a. Urgently needed: AI model evaluation from meteorological perspective. Despite the initial performance results and tremendous opportunity, a substantial amount of work

remains prior to potential transition to operations. The primary metrics used for evaluation so far are RMSE and ACC applied to all model outputs. While this is a natural and necessary starting point, many other aspects need to be urgently investigated (Ebert-Uphoff and Hilburn 2023). In particular, AI-based models behave differently from physically based models, and forecasters may need to learn the strengths and weaknesses of AI-based models to help inform their forecasts. Key topics to be investigated include the following:

- The feature-based verification of events such as tropical cyclones, atmospheric blocking, and atmospheric rivers with metrics targeted to the specific feature is a crucial next step to determine if AI-based models can accurately depict impactful meteorological phenomena.
- Similarly, the verification of performance for extremes of traditional variables is a priority. A common sentiment within the meteorological community is that AI-based models struggle to represent extreme events due to their low prevalence in training datasets (such as intense precipitation in Lavers et al. 2022) and the application of L1 (mean absolute error; Bi et al. 2023), L2 (mean-square error; Pathak et al. 2022; Bonev et al. 2023), or pressure-weighted L2 loss (Lam et al. 2023), which may discourage predictions toward the tail ends of variable distributions. Preliminary evaluation has shown AIWP models to be surprisingly skillful when it comes to these types of extremes (Lam et al. 2023; Bouallègue et al. 2024; Charlton-Perez et al. 2024), but additional independent validation on a robust reforecast dataset would be extremely valuable.
- Most AIWP models implement no physical constraints to ensure temporal consistency and consistency between variables. This has potential ramifications for the interpretability of output and the capability of forecasters to diagnose model errors. While recent work has shown that AIWP models tend to encode realistic physics (Hakim and Masanam 2023), this is not guaranteed to be the case in all situations and the identification of these potential scenarios could be a valuable contribution.
- Another critical component is gathering qualitative feedback from potential users of the model output to help document model behavior and identify model instabilities and inconsistencies that may not be evident in bulk verification.
- Last, we need to investigate AIWP with a social science lens to evaluate National Weather Service (NWS) forecaster perspectives, including the trustworthiness and operational utility of AIWP. As discussed by Murphy (1993), forecasts have no intrinsic value. They acquire value through their ability to influence the decisions made by the users of the forecasts. With respect to AIWP, this means that even if AIWP outperforms NWP, the models only have a value if weather forecasters and the public trust them and put their forecasts into practice. For example, Murphy and Ehrendorfer (1987) demonstrate that the forecast value can even decrease as forecast accuracy increases. Furthermore, it is important to engage AI users and other stakeholders in the development of AI models for environmental sciences (Bostrom et al. 2024). Early involvement of and co-development with forecasters is crucial to ensure that forecasters understand AIWP strengths and weaknesses, establish forecaster trust, and maximize the forecast value of AI-based models. Questions to be investigated in this domain include the following: Does the lack of physical underpinning reduce forecaster trust or likelihood to use a model? And is there value in efforts to operationalize AIWP models limited to 0.25° resolution and basic output variables or should we instead wait for advancements toward mesoscale features and impactful variables?

It is clear that a concerted effort by the meteorological community is urgently needed to evaluate such AI models before these models can be used in operations. However, researchers and forecasters can only evaluate what they have access to, and currently, access to the model outputs is limited.

b. Providing easy access to AIWP model output to accelerate their evaluation. Feature-based and extreme event verification, qualitative model interrogation, and evaluation of forecaster perspectives all require access to AIWP output. The ECMWF provides a major service to the community, contributing uniform code for four major AIWP models in a GitHub repository (ECMWF Lab 2023) that makes it easier to install and run these models locally than using their original code bases. However, while AIWP models are substantially cheaper to run than traditional NWP models, powerful graphics cards, or graphics processing units (GPUs), are still necessary, and significant effort is required to download and/or modify [IFS or Global Forecast System (GFS)] initial conditions to generate forecasts, placing the models out of reach for many potential users. Our own experience has shown that a GPU with at least 16 GB of memory is the minimum requirement for FourCastNet v2-small and Pangu-Weather, while 35 GB is required for GraphCast Operational.

To conserve resources due to duplicate efforts and to make AIWP model evaluation accessible to all, we present two new resources to facilitate the quantitative and qualitative investigation of AIWP models. First, we detail a 3-yr archive of FourCastNet v2-small, Pangu-Weather, and GraphCast Operational reforecasts that can be used for the robust verification of meteorological features. To the best of our knowledge such a public archive does not yet exist. Details of these three models along with the full version of GraphCast are presented in Table 1.

Second, we present an experimental web page that visualizes real-time output of the same models side-by-side with traditional, physics-based NWP models. The ECMWF also provides real-time output of several models in their AI model charts (ECMWF 2023). One key difference between these two efforts is that the ECMWF's AI model charts show AIWP forecasts based on IFS initialization, while we provide AIWP forecasts based on GFS initialization—a topic discussed further in the next section. Additionally, our visualization tool provides an interactive interface that keeps the needs of forecasters and other evaluators in mind, such as being able to easily examine specific regions around the globe and comparing results from different models with ease. This real-time data are also appended to the reforecast archive as netCDF files.

Table 1. Characteristics of FourCastNet v2-small, Pangu-Weather, and GraphCast Operational. The full GraphCast model configuration is included for completeness and its relationship to GraphCast Operational. GraphCast Operational has not been officially published and so the test period is unknown. Train refers to the dataset used to learn data patterns and relationships, validation to the dataset used to tune model parameters, and test to the independent dataset used to evaluate model skill. *u* = *U*-wind component; *v* = *V*-wind component; *T* = temperature; *z* = geopotential heights; *r* = relative humidity; *q* = specific humidity; *w* = vertical velocity; TCWV = total column-integrated water vapor.

	Architecture	Train	Validation	Test	Archive availability	Surface variables (inputs/outputs)	Pressure level variables (inputs/outputs)
AI model configurations							
FourCastNet v2-small	Spherical Fourier neural operators	1979–2015	2016–17	2018	October 2020–the present	u10, v10, u100, v100, T2, MSLP, SP, TCWV	*u*, *v*, *t*, *z*, *r* @ 1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 50 hPa
Pangu-Weather	3D Earth specific transformer	1979–2017	2019	2018	October 2020–the present	u10, v10, T2, MSLP	*u*, *v*, *t*, *z*, *q* @ 1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 50 hPa
GraphCast	Graph neural network	1979–2015	2016–17	2018–21	—	u10, v10, T2, MSLP, APCP[a]	*u*, *v*, *t*, *z*, *q*, *w* @ 1000, 975, 950, 925, 900, 875, 850, 825, 800, 775, 750, 700, 650, 600, 550, 500, 450, 400, 350, 300, 250, 225, 200, 175, 150, 125, 100, 70, 50, 30, 20, 10, 7, 5, 3, 2, 1 hPa
GraphCast Operational	Graph neural network	1979–2017	fine-tuned 2016–21	—	January 2022–the present	u10, v10, T2, MSLP, APCP[a]	*u*, *v*, *t*, *z*, *q* @ 1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 50 hPa

[a] APCP is an output but not an input for GraphCast Operational.

2. Initial conditions—Using the GFS versus the IFS

Given the fact that the models presented in Table 1 were trained on ERA5 reanalysis, it was a natural decision for the developers of the models to both verify performance against the ERA5 and develop real-time applications with the IFS initial conditions. Motivated by real-time applications at the ECMWF, the GraphCast team even went so far as to fine-tune an operational version of their model for improved performance with HRES IFS initial conditions. Extending from reanalysis to analysis resulted in only a marginal decrease in forecast skill as measured by RMSE due to the strong correlation between ERA5 and IFS analyses (Lam et al. 2023).

Though there are clear advantages to running AIWP models on the IFS initial conditions, there is also potential justification for running on the GFS initial conditions. For example, for us and others in the AI development community, the necessary real-time IFS conditions to initialize AIWP models only just became available. Furthermore, there may be a preference within the NOAA community toward initialization with a product internal to NOAA (like the GFS) due to greater familiarity. Finally, it is still an open question how extensible AIWP is to unfamiliar initial condition datasets. For the above reasons, we provide reforecasts initialized with both the IFS and GFS initial conditions. Our real-time implementation/visualization web page of AIWP models is currently only initialized with the GFS, with the understanding that performance is likely not maximized with this configuration. We hope to expand this to include the IFS initialized versions in the near future. For full clarity, we initialize all models on forecast hour zero (F000) data at 0.25° grid spacing, rather than any analysis products.

The variables necessary for initialization of FourCastNet v2-small, Pangu-Weather, and GraphCast Operational are generic enough that they are almost all present in most initial condition datasets, including the GFS. This generally includes geopotential, temperatures, specific humidity, and u- and v-wind components at 13 vertical levels, in addition to mean sea level pressure (MSLP), 2-m temperature, and 10-m u- and v-wind components (see Table 1 for full lists). Geopotential was calculated from the GFS's geopotential height fields. The only variable necessary that is not natively present in the GFS files is top of atmosphere incident solar radiation (TOA radiation). TOA radiation is a function solely of date/time, latitude, and longitude, and thus, we felt justified in using ERA5 values rather than calculate it ourselves. For real-time output, GraphCast developers have recently provided a Python library for TOA radiation calculation (Google DeepMind 2024), which will be applied moving forward.

3. Reforecast archive

The reforecast archive includes FourCastNet v2-small, Pangu-Weather, and GraphCast Operational and generally dates back to October 2020 with two initializations per day (0000 and 1200 UTC). The GraphCast Operational archive only goes back to January of 2022 so as not to overlap with the "fine-tuning" period. Forecasts are run out to 240 h at 6-h time steps. This period was chosen for two reasons. First, it is 3 years retrospective from the date that we began to develop the archive. Second, the period is independent from the training and validation periods used to develop the three models. We may consider extending the FourCastNet v2-small and Pangu-Weather archive further back due to their earlier training and validation periods.

The archive is provided through the NOAA's Open Data Dissemination (NODD) via an Amazon Simple Storage Service (S3) bucket (NOAA-OAR 2024). Each model initialization (all forecast steps) is stored in a netCDF4 file with level four zlib compression. Each FourCastNet v2-small, Pangu-Weather, and GraphCast Operational file is approximately 7–9 GB. The full archive is accessible online (https://noaa-oar-mlwp-data.s3.amazonaws.com/index.html).

All variables output by the individual models are included in the files. These vary slightly by model, but include geopotential, temperature, and U- and V-wind components at 1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, and 50 hPa for all three models (Bonev et al. 2023; Bi et al. 2023; Lam et al. 2023). Pangu-Weather and GraphCast Operational

contain specific humidity at these levels (Bi et al. 2023; Lam et al. 2023), while FourCastNet v2-small has relative humidity (Bonev et al. 2023). GraphCast Operational also has vertical velocity at these levels (Lam et al. 2023). In terms of surface variables, all three have 2-m temperature, mean sea level pressure, and 10-m *U*- and *V*-wind components (Bonev et al. 2023; Bi et al. 2023; Lam et al. 2023). FourCastNet v2-small also contains 100-m *U*- and *V*-wind components, surface pressure (SP), and precipitable water (Bonev et al. 2023), while GraphCast Operational adds 6-hourly accumulated precipitation (APCP) (Lam et al. 2023). The full list can be seen in Table 1.

Real-time AIWP model output is now being appended to continuously grow the archive. Additionally, promising new models will be added to the archive both in retrospect and in real-time as they are released. How new models are chosen for inclusion is an inherently subjective process. We considered applying minimum performance criteria, but this is also subjective and potentially excludes models that outperform in alternative metrics or applications. Rather, we list here some of our primary considerations for inclusion:

- In order for us to be able to run new models, they must be open access.
- Though we define no minimum performance criteria, new models should make advances in the prediction of meteorological phenomena for some metrics or applications.
- To avoid repetition in the archive, preference will be given to models with novel AI/machine learning (ML) approaches.
- GPU resources are limited and we may be unable to run models that have substantial memory requirements. Thus, we encourage developers to strongly consider limiting model size, whenever possible.
- Preference may also be given to models developed by organizations with proven weather forecasting track records.

At present, this archive provides only deterministic AIWP model forecasts. The generation of well-calibrated ensembles or probabilistic AIWP model forecasts is another exciting new research topic and is generally not as simple as perturbing the initial conditions of deterministic AIWP models (Selz and Craig 2023; Price et al. 2024). Producing sufficient spread (Selz and Craig 2023) and maintaining sharpness in individual member solutions are particularly challenging tasks for which Price et al. (2024) have made promising progress. However, given that AIWP ensembles are in early stages, are computationally expensive, and can potentially produce orders of magnitude more data than deterministic AIWP models, inclusion is currently outside the scope of this project. We may consider adaptations to the archive to allow for probabilistic data in the future.

4. Real-time output and visualizations

All of the variable information on the reforecast archive is also applicable to real-time implementation of the models, with output appended to the S3 bucket noted above. However, rather than requiring potential users to download and process the data themselves to view the output and familiarize themselves with AIWP models, we have prepared an experimental web page (Radford 2023) that visualizes all variables for the four most recent initializations of all three models. Like Radford et al. (2023a,b), the data are displayed on an interactive map to facilitate panning and zooming to regions and scales of interest and increase customizability. Hover readout capabilities to view values at a point have also been implemented. Data are currently visualized using vectors to allow for future customizability of contours and colormaps.

The primary goal of this web page is to allow users to familiarize themselves with the output of AI models and determine whether the output aligns with expectations and/or is physically reasonable. An additional goal is to facilitate intermodel comparisons, not just between

different AIWP models but also between AIWP models and traditional, dynamical models. Toward this end, we also visualize the output of the GFS and the IFS. Switching between variables or between models keeps every other parameter unchanged to make comparisons easier. The four most recent initializations can be viewed to evaluate model solution evolution over time. Last, we offer a four-panel comparison tool that can be used to view four different models side-by-side as in Fig. 1. Panning and zooming in one map can be mirrored by the others to keep the views locked to the same area.

This is an experimental tool for the community to build familiarity with and assess the trustworthiness of AIWP. It is not operational and thus does not have 24/7 technical support and is likely to experience unexpected downtime. That said, we hope to continue developing the page to add features and new models. We would encourage users to use the "contact us" page to make requests for features, identify bugs, and note observations of AIWP model behavior and performance. We will also be adding new AIWP models as they become available and according to the considerations listed in the reforecast archive section.

We have tested this tool in daily weather briefings at CIRA and found that it has refined our conceptions of AI model behavior, strengths, and weaknesses. We also demonstrated this tool to several forecasters, e.g., at the NOAA AIWP workshop (NOAA-PSL 2023), and received a very enthusiastic response from the forecasters, i.e., they were very interested in giving it a try. With minimal publicization, the web page currently receives approximately 1000 unique visitors per month, indicating strong interest within the meteorological community.

5. Discussion

AIWP models are powerful new forecasting tools that will likely have an increasing influence in the operational forecast community over the coming years. It is crucial that we begin to familiarize the community with the output of these models in both quantitative and qualitative frameworks so that users can gauge the trustworthiness and utility of AIWP.

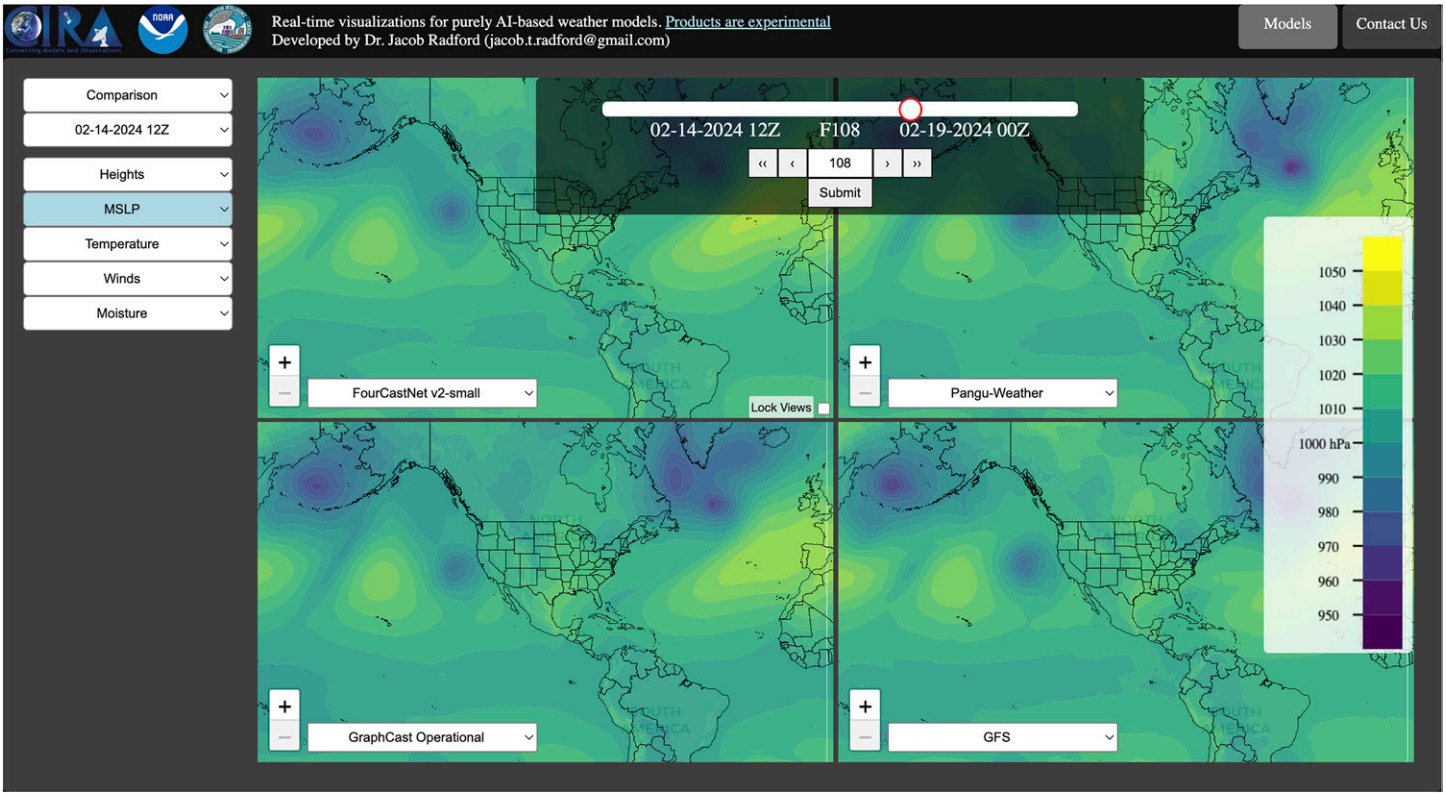

Fig. 1. Screenshot of the four-panel comparison tool on the AIWP visualizations web page viewing MSLP for 1200 UTC 14 Feb 2024 initializations of FourCastNet v2-small, Pangu-Weather, GraphCast Operational, and the GFS.

We now provide two resources to start progress in this direction. First, our reforecast archive allows users with various interests and expertise to perform evaluations of AIWP forecast quality for specific meteorological features or extreme events, at least to the extent that the necessary variables are present in the output files. For example, features that we believe would be valuable to investigate include tropical cyclones (such as in DeMaria 2024), fronts, atmospheric rivers, and severe storm parameters. Other interesting phenomena will have to wait until additional output fields like hydrometeor mixing ratios and wind gusts are made available, but our goal is to continue to update our archive with the latest models and versions to facilitate this future work.

Second, our web-based visualization page allows interested users to investigate AIWP model output and compare that output to models with which they are already familiar. Users can then begin to apply their own judgment on whether the output of AIWP models is physically reasonable, trustworthy, and useful. Both resources could help to foster communication between the development and operational communities on how to best improve AIWP models moving forward, a key relationship that, to this point, has often been neglected. For example, the National Severe Storms Laboratory (NSSL) utilized the reforecast archive in their 2024 annual Spring Forecasting Experiment (SFE) to collect forecaster feedback on AIWP forecast value in severe storms forecasting. We anticipate more testbeds will follow suit and begin exploring the value of AIWP in different operational contexts. We also strongly encourage users of the visualizations page to use the feedback form to document AIWP model successes, failures, and oddities, as well as request additional features or resources which we can pass on to the respective model developers.

It is our opinion that AIWP models have a bright future and will play an important role in weather forecasting over the next several years. However, we want to note that at this point, we are not pushing for operationalizing AIWP models. By providing the reforecast archive and visualization web page, we merely want to help facilitate the rigorous evaluation necessary before considering use in operations and ease the transition for users if operationalizing is ultimately warranted.

Acknowledgments. Author Radford is supported by funding from NOAA Award NA19OAR4320073. Ebert-Uphoff's work on this topic is supported by CIRA's ML strategic initiative. The authors thank the developers of FourCastNet, Pangu-Weather, and GraphCast for publishing their model code as well as the ECMWF for providing the ai-models package to simplify running these models, as open access code and data are of paramount importance toward advancing the AIWP field. Last, we thank the NOAA Office of Oceanic and Atmospheric Research (OAR) for providing the AWS S3 bucket.

Data availability statement. The authors used a local archive of 0.25° GFS initial conditions. Those data are also available via Amazon Web Services (AWS; https://registry.opendata.aws/noaa-gfs-bdp-pds/). The model codes and ai-models package used in this study are available on GitHub: FourCastNet (https://github.com/NVlabs/FourCastNet), Pangu-Weather (https://github.com/198808xc/Pangu-Weather), GraphCast (https://github.com/google-deepmind/graphcast), and AI models (https://github.com/ecmwf-lab/ai-models). The archive described in the paper is now available at https://noaa-oar-mlwp-data.s3.amazonaws.com/index.html. The visualization web page is accessible at https://aiweather.cira.colostate.edu.

References

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

Bonev, B., T. Kurth, C. Hundt, J. Pathak, M. Baust, K. Kashinath, and A. Anandkumar, 2023: Spherical Fourier neural operators: Learning stable dynamics on the sphere. arXiv, 2306.03838v1, https://doi.org/10.48550/arXiv.2306.03838.

Bostrom, A., and Coauthors, 2024: Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences. *Risk Anal.*, **44**, 1498–1513, https://doi.org/10.1111/risa.14245.

Bouallègue, Z. B., and Coauthors, 2024: The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bull. Amer. Meteor. Soc.*, **105**, E864–E883, https://doi.org/10.1175/BAMS-D-23-0162.1.

Charlton-Perez, A. J., and Coauthors, 2024: Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of storm Ciarán. *npj Climate Atmos. Sci.*, **7**, 93, https://doi.org/10.1038/s41612-024-00638-w.

DeMaria, R. T., M. DeMaria, G. Chirokova, K. Musgrave, J. T. Radford, and I. Ebert-Uphoff, 2024: Evaluation of tropical cyclone track and intensity forecasts from purely ML-based weather prediction models, illustrated with FourCastNet. *23rd Conf. on Artificial Intelligence for Environmental Science*, Baltimore, MD, Amer. Meteor. Soc., 4A.2, https://ams.confex.com/ams/104ANNUAL/meetingapp.cgi/Paper/436711.

Ebert-Uphoff, I., and K. Hilburn, 2023: The outlook for AI weather prediction. *Nature*, **619**, 473–474, https://doi.org/10.1038/d41586-023-02084-9.

ECMWF, 2023: Products from various AI models. Accessed 11 February 2024, https://charts.ecmwf.int.

ECMWF Lab, 2023: AI-models. Accessed 11 February 2024, https://github.com/ecmwf-lab/ai-models.

Google DeepMind, 2024: Graphcast. Commit e70b8ca, January 3, 2024. Accessed 28 August 2024, https://github.com/google-deepmind/graphcast/commit/e70b8cae886777197262ee2b01585d5ef5065442.

Hakim, G. J., and S. Masanam, 2023: Dynamical tests of a deep-learning weather prediction model. arXiv, 2309.10867v1, https://doi.org/10.48550/arXiv.2309.10867.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Lam, R., and Coauthors, 2023: Learning skillful medium-range global weather forecasting. *Science*, **382**, 1416–1421, https://doi.org/10.1126/science.adi2336.

Lang, S., and Coauthors, 2024: AIFS—ECMWF's data-driven forecasting system. arXiv, 2406.01465v2, https://doi.org/10.48550/arXiv.2406.01465.

Lavers, D. A., A. Simmons, F. Vamborg, and M. J. Rodwell, 2022: An evaluation of ERA5 precipitation for climate monitoring. *Quart. J. Roy. Meteor. Soc.*, **148**, 3152–3165, https://doi.org/10.1002/qj.4351.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

——, and M. Ehrendorfer, 1987: On the relationship between the accuracy and value of forecasts in the cost–loss ratio situation. *Wea. Forecasting*, **2**, 243–251, https://doi.org/10.1175/1520-0434(1987)002<0243:OTRBTA>2.0.CO;2.

NOAA-OAR, 2024: NOAA-OAR-MLWP-data. Accessed 11 February 2024, https://noaa-oar-mlwp-data.s3.amazonaws.com/index.html.

NOAA-PSL, 2023: Artificial Intelligence for Numerical Weather Prediction (AI4NWP) workshop. Accessed 11 February 2024, https://psl.noaa.gov/events/2023/ai4nwp_workshop/.

Pathak, J., and Coauthors, 2022: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv, 2202.11214v1, https://doi.org/10.48550/arXiv.2202.11214.

Price, I., and Coauthors, 2024: GenCast: Diffusion-based ensemble forecasting for medium-range weather. arXiv, 2312.15796v2, https://doi.org/10.48550/arXiv.2312.15796.

Radford, J., 2023: Real-time visualizations for purely AI-based weather models. Accessed 11 February 2024, https://aiweather.cira.colostate.edu.

Radford, J. T., G. M. Lackmann, J. Goodwin, J. Correia, and K. Harnos, 2023a: An iterative approach toward development of ensemble visualization techniques for high-impact winter weather hazards: Part I: Product development. *Bull. Amer. Meteor. Soc.*, **104**, E1630–E1648, https://doi.org/10.1175/BAMS-D-22-0192.1.

——, ——, ——, ——, and ——, 2023b: An iterative approach toward development of ensemble visualization techniques for high-impact winter weather hazards: Part II: Product evaluation. *Bull. Amer. Meteor. Soc.*, **104**, E1649–E1669, https://doi.org/10.1175/BAMS-D-22-0193.1.

Rasp, S., and Coauthors, 2024: Weatherbench 2: A benchmark for the next generation of data-driven global weather models. arXiv, 2308.15560v2, https://doi.org/10.48550/arXiv.2308.15560.

Selz, T., and G. C. Craig, 2023: Can artificial intelligence-based weather prediction models simulate the butterfly effect? *Geophys. Res. Lett.*, **50**, e2023GL105747, https://doi.org/10.1029/2023GL105747.