Global reforecasts from MPAS GRAF with mesh refinement over the US and Europe

Thomas M. Hamill[1], Raghu Raj Prasanna Kumar[2], Karthik Kashinath[2], Carl Ponder[2], Mike Pritchard[2], Tao Ge[2], Akshay Subramanian[2], Jaideep Pathak[2], John Wong[1], Brett Wilt[1], and Peter Neilley[1]

[1] *The Weather Company (TWC), Atlanta, Georgia USA.*

[2] *NVIDIA, Santa Clara, California, USA.*

Abstract

NVIDIA and The Weather Company (TWCo) have generated a data set of reforecasts from TWCo's GRAF (Global high-Resolution Atmospheric Forecasting) model, a version of the National Center for Atmospheric Research (NCAR) Model for Predictions Across Scales (MPAS). GRAF is global, but the configuration for this reforecast had a mesh refinement to ~4 km over the US, Caribbean Basin, and Europe, and 15 km elsewhere. This model was designed to run much of the computation on graphical processing units, with this development assisted by NVIDIA.

The 1836 reforecast cases (~5 years) were generated from ECMWF reanalyses (ERA5) for selected initial condition dates spanning more than 20 years, 2004-2024. These dates of the chosen initial conditions mostly selected based on high-impact weather in the contiguous US (CONUS) and Caribbean. Sampling in this way, the reforecast spanned a wider range of interesting, high-impact weather scenarios than had we performed five contiguous years of once-daily reforecasts. The reforecast still provides many samples in non-precipitating regions with more ordinary weather.

GRAF reforecasts were mostly run to +27 h lead time, assuming a 3-h for spin up followed by a full diurnal cycle. Data were saved in zarr format on the native model vertical coordinate. Most fields were archived at 15-min intervals, though several precipitation variables were saved at 5-min cadence. Data are made publicly available to all through Amazon Web Services' Open-Data Initiative.

1. Introduction

Many organizations in the weather enterprise seek to provide ever-improving predictions, from which customers can make better and better decisions. Numerical weather prediction (NWP) is the underpinning technology behind accurate predictions of weather changes. These improvements have accumulated at a rate of about one day a decade, i.e., a four-day forecast now is as accurate as a three-day forecast produced a decade ago. The slow accumulation of skill, which translates into improved products and services, represents a “quiet revolution” in weather prediction (Bauer et al. 2015).

Within the last few years and with the advance of artificial intelligence, a radically different approach to NWP has been developed. The new models are *data driven*; they do not represent a complex human codification of the physical laws of motion, parameterized processes, and the interactions between state components. Instead, comparatively simple neural-network models are trained. In a common method of coding these models for weather prediction, this provides a sophisticated mapping from the current atmospheric state to the state a few hours or days hence; the weights used in the neural network are chosen to minimize error, often root-mean square or mean absolute error. These mappings are typically chained together to provide a prediction; from the current state, a forecast is made to six hours in the future; from the forecast at six hours, a forecast is made to twelve hours hence, and so forth. Henceforth we will refer to these data-driven models as deep-learning NWP models, or DLNWP. The complexity of these models is substantial but is hidden within the neural network; the actual number of lines of code tailored to the prediction application is very small compared to conventional NWP, perhaps by an order of 100 or more.

The first low-resolution, proof-of-concept DLNWP models were developed only in the late 2010’s (Deuben et al. 2018) and were significantly less accurate than conventional NWP forecasts. Informed by these early test systems, more computational horsepower was made available to train more sophisticated DLNWP models, and these have increased in skill at a dramatic rate. Advanced DLNWP development was demonstrated in Weyn et al. (2020, 2021), GraphCast (Lam et al. 2023), Met-Net-3 (Andrychowicz et al. 2023), NeuralGCM (Kochkov et al. 2024), GenCast (Price et al. 2024) FourCastNet (Pathak et al. 2022) and CorrDiff (Mardani et al. 2023), ClimaX (Nguyen et al. 2023), Aurora (Bodnar et al. 2024), FengWu (Han et al. 2024), FuXi (Chen et al. 2023) and many more. DLNWP represents a radical change for weather prediction – simplified code that bypasses the many complex parameterizations of physical processes and yet may produce more accurate forecasts. In selected ways of measuring weather forecast skill, several of these are now competitive with or more skillful than raw conventional numerical forecasts from the world-leading European Centre for Medium-Range Weather Forecasts (ECMWF).

A particular limitation that motivated the generation of this data set is the comparative paucity of demonstrated skillful predictions at the scale of phenomena that are of concern to weather users. Many of these users want actionable detail, if possible, on high impact weather that can occur at scales down to a few km. The ability of DLNWP developers to begin

providing such products may be facilitated by high-resolution training data sets such as convection permitting reanalysis downscaling (e.g., “CONUS404” , Rasmussen et al. 2023) or reforecasts. These may provide information on inferring the small-scale details that are consistent with the larger-scale meteorological forcings provided by the successful global DLNWP systems. While reanalyses (analysis states produced through cycled data assimilation) are conceptually preferable to reforecasts (retrospective forecast states), reanalyses are typically much, much more computationally demanding and require more thought and care, e.g., Hersbach et al. (2020), Hamill et al. (2021). Hence, we did not generate a reanalysis ourselves in this project. We note that another data set users may consider, CONUS404, isn’t a high-resolution reanalysis itself but rather is a fine-resolution model downscaling nudged to a larger-scale reanalysis. Some convective-scale deep learning algorithms may require data more frequently than CONUS404’s hourly output as well.

To jumpstart convection-permitting DLNWP, we thus decided to generate a convection-permitting reforecast data set with mesh refinement to 4 km over much of North America and the Caribbean, as well as western Europe (Fig. 1). Most data are saved at 15-min cadence. Our companies, NVIVIA and The Weather Company, have subsequently developed deep learning numerical weather prediction systems with these data. We now provide the data without cost to enterprise partners to facilitate others’ ability to advance DLNWP.

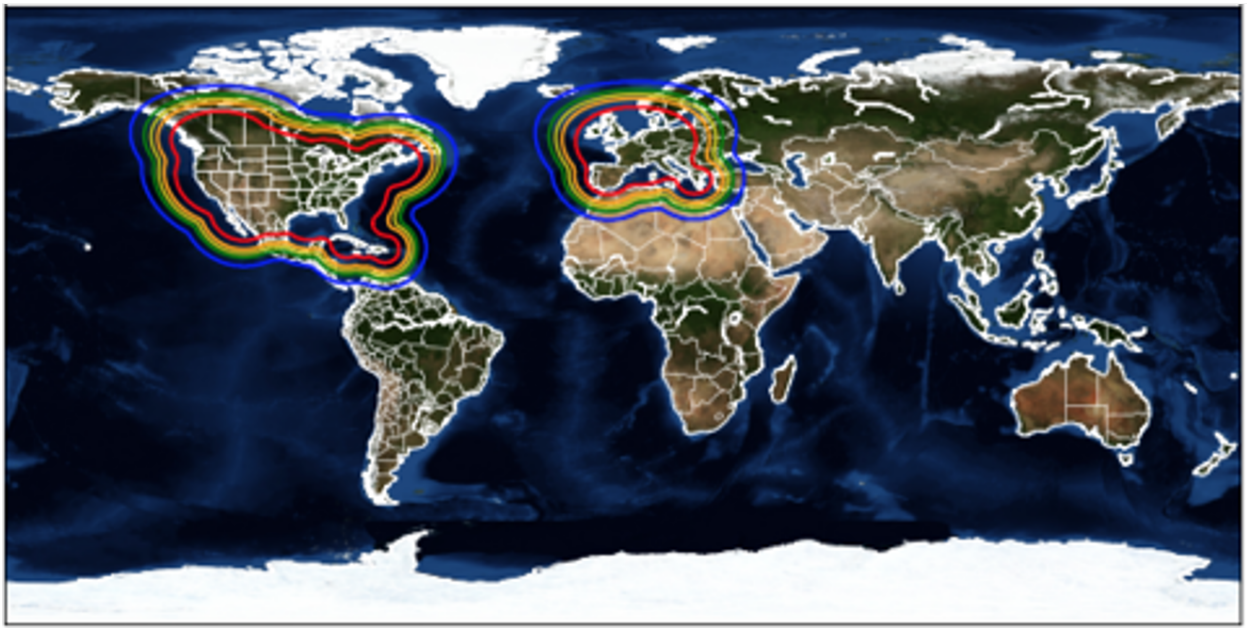

Figure 1: Illustration of the variable grid spacing for the global MPAS model to be used in the reforecasts described here. Grid spacing is ~4 km inside red the line, ~15 km outside the blue area.

The rest of the document describes model configuration details, the procedure that was used to define the dates of initial conditions, and information on the data set storage.

2. Model configuration.

The TWC ”GRAF” (Global high-Resolution Atmospheric Forecasting) model is based on the NCAR MPAS (Model for Prediction Across Scales; Skamarock et al. 2012, Park et al. (2014), Sandbach et al. (2015), Heinzeller et al. (2016)), version 6.3. Basic details are included in Table 1, and zeta values used in the determination of 50 model-level heights are provided in Table 2. The model has a ~ 4-km grid spacing over the CONUS and Europe, relaxing to 15 km elsewhere (Fig. 1). It employs a scale-aware nTiedtke convective parameterization (Wang 2022) with near full use of the convective parameterization at 15 km to near none at 4 km. It also uses the YSU planetary boundary layer and gravity-wave drag, WSM6 microphysics (Hong and Lim 2006) with Thompson cloud fraction, RRTMG radiative transfer scheme (Mlawer et al 1997, Clough et al. 2005), and the NOAH 4-layer land-surface model (Ek et al. 2003). The terrain “GMTED2010/MODIS 30 arcsec” (Global Multi-resolution Terrain Elevation Data (https://www.usgs.gov/coastal-changes-and-impacts/gmted2010) and albedo and land characteristic data from MODIS 30 arcsec (https://modis-land.gsfc.nasa.gov/brdf.html). The model was initialized from ERA5 reanalyses on model levels (Hersbach et al 2020).

Model	MPAS 6.3 (using GPUs, with TWCo tuning), 50 levels, ~ 4.8 M grid cells
Physics	WSM6 cloud microphysics; nTiedtke scale-aware convective parameterization; YSU planetary boundary layer and gravity wave drag; surface layer via Monin-Obukhov option. RRTMG radiative transfer scheme; NOAH land-surface model; Thompson cloud fraction
Initialization	MPAS 7.0
Terrain	Global Multi-resolution Terrain Elevation Data 2010
Albedo	MODIS Land, 30 arcsec grid spacing.
Land Use	MODIS-derived predominant land type, veg fraction.
Atmospheric state, sea-surface temperature, soil state	ERA5 reanalysis on model levels, 0.25-deg.

Table 1: Configuration for the GRAF reforecast vs. the operational GRAF system at TWCo.

Model level	Zeta (m)	Model level	Zeta (m)	Model level	Zeta (m)
1	0	18	7528	35	24256
2	65.9983	19	8512	36	25240
3	146.143	20	9496	37	26224
4	248.677	21	10480	38	27208
5	379.614	22	11464	39	28192
6	545.524	23	12448	40	29176
7	752.977	24	13432	41	30160
8	1008.55	25	14416	42	31144
9	1318.8	26	15400	43	32128
10	1690.31	27	16384	44	33112
11	2129.64	28	17368	45	34096
12	2643.37	29	18352	46	35080
13	3238.07	30	19336	47	36064
14	3920.31	31	20320	48	37048
15	4696.65	32	21304	49	38032
16	5573.68	33	22288	50	39016
17	6544	34	23272	51	40000

Table 2: A list of the zeta values used in conjunction with terrain elevation to determine the model coordinate heights.

3. Approach for choosing the dates of the initial conditions.

The initial conditions dates were selected in multiple ways but generally prioritized dates expected to have high-impact weather in the contiguous US. These included choosing cases for near-landfalling US/Mexican hurricanes, US severe local storms, and dates with previously forecast heavy precipitation in major US hydrologic units. A small number of cases were selected at the end of the process, filling in the largest gaps in sequences of dates for weather-dependent reforecasts.

a. *Selecting the dates for tropical cyclones, severe local storms, and other high-impact weather phenomena.*

This part of the initial-condition date selection process was not fully objective in character. Wikipedia pages for US Atlantic hurricane seasons from 2004-2023 were examined (e.g., here), and initial dates were chosen typically 12-24 h preceding US or near-US land-falling hurricanes. Occasionally cases were chosen for US tropical storms as well. With hurricanes that lingered over land such as Harvey in 2017 (Houston-area floods), or with hurricanes that stayed near the coast such as 2019's Dorian, multiple reforecast dates were chosen closely spaced in time and covered more than just landfall. The only eastern Pacific hurricane date was for Otis (2023), which rapidly intensified unexpectedly and made landfall at Acapulco, Mexico. Given the reforecasts extend to only a short lead time, we do not envision this method as providing, in itself, a quantification of track or intensity skill beyond the 27 h of the reforecast. Hurricane Sandy (2012) was accidentally omitted.

Many case dates were also selected based on the criteria of either observed or forecast severe weather and tornado outbreaks. Again, wikipedia pages were examined for severe local storms (e.g., here), and dates with many tornadoes or a few significant tornadoes were chosen. Since most tornadoes occur in late afternoon, typically a 12 UTC initialization time was commonly chosen. A few cases with overnight tornadoes were initialized at 18 or 00 UTC.

We also examined the literature for dates of predecessor rain events (PREs, Galarneau et al. 2010) associated with tropical cyclones, major northeast US snowstorms, ice storms, and major west-coast atmospheric rivers (Zhu and Newell 1994), and we included a few dates for these. These lists are admittedly incomplete; the list of PREs was based on dates listed in a 2010 journal article, so all cases precede 2010. Northeast US snowstorms were based on Google searches and the memories of TWCo employees. We wanted to find dates of major mesoscale convective complexes, as these are often poorly modeled in numerical guidance and occur in weaker synoptic flow regimes; we found no online references for these, regrettably.

A synthesis of the cases selected based on these criteria are shown in Fig. 2.

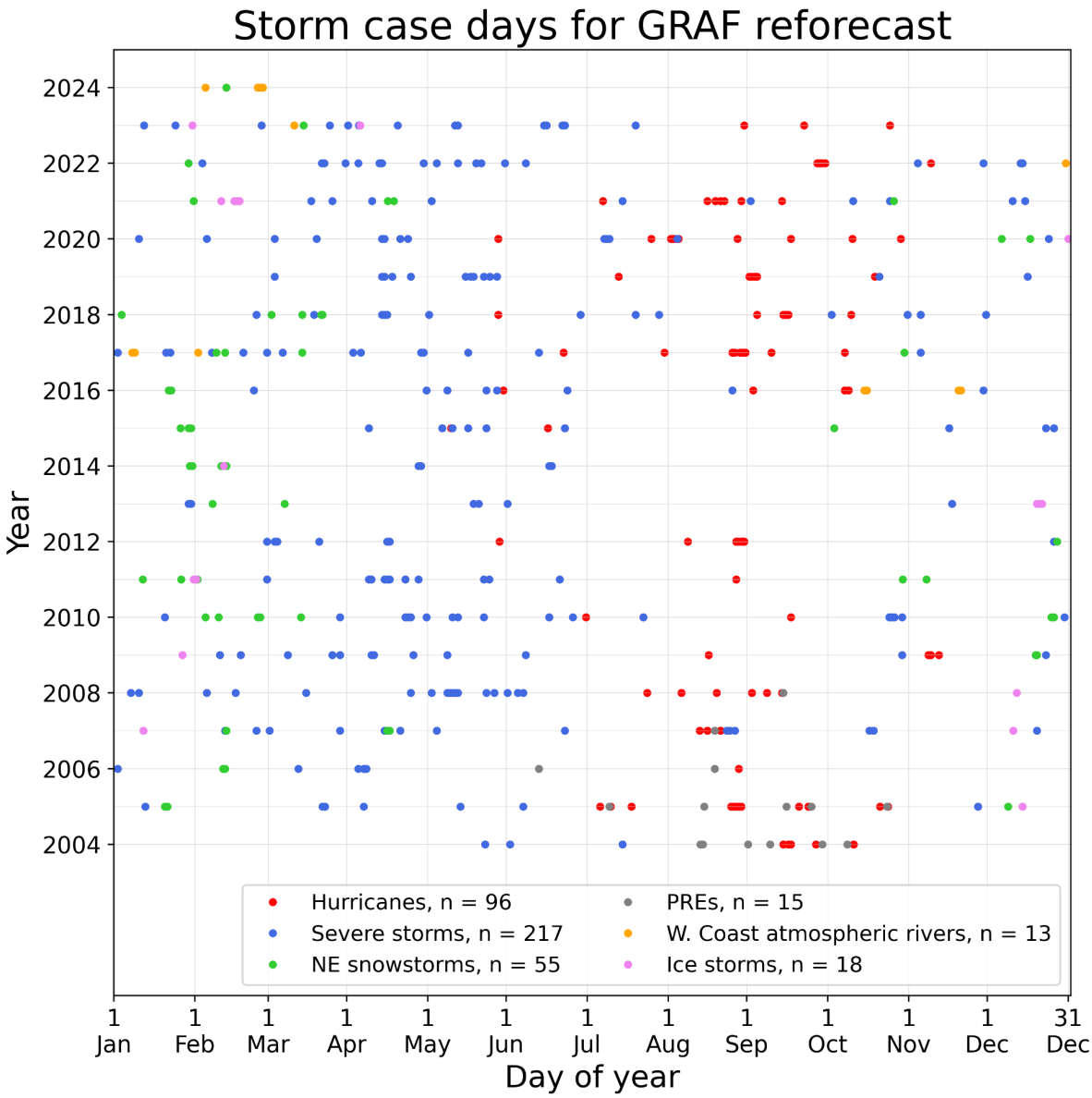

Figure 2: Illustration of the 396 dates chosen based for severe local storms, hurricanes, northeast US snowstorms, predecessor rain events, and west-coast atmospheric rivers.

b. *Selecting most of the cases based on heavy forecast rainfall in major hydrologic units.*

Since the primary initial use case for the TWCo-NVIDIA collaboration is to develop and demonstrate a capacity for high-resolution, deep-learning based ensemble forecasts of precipitation, most of the remaining cases were selected based on heavy forecast ensemble-mean precipitation somewhere in the CONUS. This provides us with a sample that has more heavy precipitation forecasts than normal, and presuming a positive forecast-to-observation correlation, more heavy observed forecasts than normal. Will this bias the machine learning training? Maybe so, but if we wish to be able to train the deep learning algorithm to provide

reasonable results spanning the most dry to the most wet scenarios, we will need sufficient samples of both wet and dry forecasts, ideally with wet samples across the domain. The assumption underlying the selection of a subset of wet forecast case dates for each major river basin is that those case dates will often have dry forecasts in regions outside that river basin. Thus, as we describe here, partitioning the CONUS into 18 separate river basins and one CONUS-wide area and finding wet cases independently for each will still result in a wide range of precipitation scenarios across the set of reforecast cases; a wet reforecast case day for a western US river basin may be a very dry day for an eastern US river basin.

To ensure we have heavy precipitation events represented across the US, we chose a subset of heavy precipitation cases for each major CONUS hydrologic unit, shown in Fig. 3 below. These "HUC-2" units were defined by the US Geological Survey (Seaber et al. 1987). Additionally, we considered precipitation averaged over all HUC-2 units.

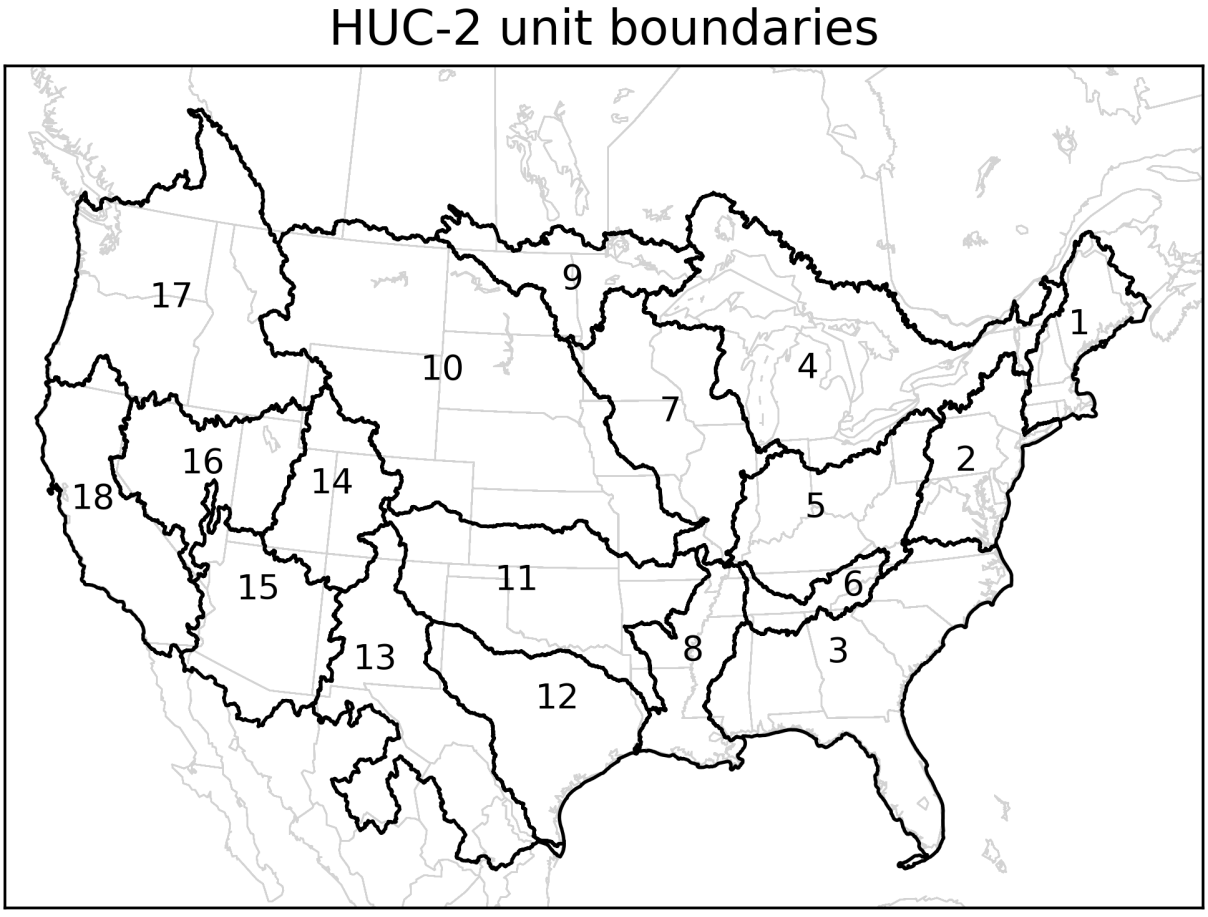

Figure 3: Illustration of HUC-2 units for the CONUS for which heavy precipitation cases were chosen. There are 18 of these, identified as numbered. The methodology also chose cases based on a 19th area, all 18 HUCs together.

We confined the reforecasts initial conditions to be primarily in the 20-year period of 2004-2023, with a few cases in early 2024; choosing initial condition dates further into the past

has the disadvantage of a being initialized with data from a thinner observation network; forecasts from the initial conditions in 1990 would thus have a lower statistical quality and higher errors than forecasts from initial conditions in recent decades. There is also the complication of the changing climate, with more extremes of precipitation in recent decades.

We chose initially to generate 715 HUC-2 case dates for both the cool season (October-March) and warm season (April-September). For each season, we divide those cases with 35 for each HUC-2 unit and the remainder, 85 samples, for the overall CONUS.

The algorithm for selecting those case dates is as follows, repeated independently for each HUC-2 unit. First, in order to avoid biasing the cases too much toward heavy *observed* precipitation, these cases are selected based on heavy *forecast* precipitation. GEFS v12 total-precipitation reforecasts for 2004-2019 (Guan et al. 2022) were used as the input data for case selection in these years. All the GEFS reforecasts were generated from 00 UTC initial conditions. GEFS v12 forecasts and reforecasts have a notable issue with spin-up, i.e., a change in the statistical character of the forecast in the forecasts' early hours. In the case of GEFS v12, there is typically anomalously heavy precipitation in the first 12 hours. Hence, as a way of quantifying forecast precipitation intensity, we determined, spatially averaged over each HUC-2 region, the 12 - 36 h ensemble-mean precipitation, a 24-h period after the worst of the spin-up should have taken place.

GEFS v12 reforecasts extend to the end of 2019, while the reforecast cases were desired for 2020-2024 as well. For these other years, TWCo used internal data. We maintain a database of GEFS ensemble forecasts coincident with regularly reporting observation stations. For similarity to the procedure above, we determined the 12-36-h ensemble-mean forecast, averaged over all the stations within each HUC-2 unit.

In general, for most HUC-2 units, there are fewer stations for the 2020-2023 data than there were model grid points for the 2004-2019 reforecasts. Hence, from central limit theorem arguments, we would expect a higher variance from the 2020-2023 station-based data. To make the statistics more uniform across the full 2004-2023 period before selecting case dates, we standardize the 2020-2023 data and then re-express with the mean and variance of the 2004-2019 data. As we process either the warm or cool season, let $\underline{x}_r$ represent the mean of the all that season's daily 2004-2019 reforecast samples for a HUC-2 unit, and let σ_r be the climatological standard deviation of those daily samples. Similarly, we compute $\underline{x}_s$and σ_s, the mean and standard deviation based on 2020-2023 GEFS ensemble-mean forecasts at observation locations. Letting x_i be the GEFS station-based ensemble-mean forecast for the i_{th} case day in the 2020-2023 period, $\hat{x}_i$ is the re-expressed mean, computed as

$$\hat{x}_i = \underline{x}_r + \sigma_r \frac{x_i - \underline{x}_s}{\sigma_s}$$

For 2020-2023, these re-expressed means then replace the raw ensemble- and station-mean averages in the time series for each basin.

The cases were then selected based largely on a rank ordering of ensemble-mean precipitation. Case dates were ordered from lowest to highest ensemble-mean precipitation, and they were assigned an initial weight based on that date's ensemble-mean precipitation, divided by the ensemble-mean precipitation for the date with the largest value. The first case date selected was the date with the largest precipitation, i.e., with the largest weight. Then, with one exception, the second and third dates have the second and third largest precipitation, and so forth through the samples with progressively less precipitation. The one exception is that in order to slightly de-emphasize the selection of dates that may be very close in time so that we have more independent samples, if a particular initial condition date was selected, the nearby dates were de-weighted so they were less likely to be chosen. For the day before, two days before, the day after, and two days after, the initial weights as described above were multiplied by a factor of 0.7. This would thus not totally eliminate such dates from consideration but gave them less probability of being chosen. The weight factor value of 0.7 was somewhat arbitrary; the rationale was that should a very major precipitation event span multiple days, we would still want a higher probability of choosing it. Figure 4 shows the HUC-2 case dates selected with the process described above.

At this point, there was no combination yet with the other cases chosen for hurricanes, severe local storms, snowstorms, and such. When we merged these data, there were some overlapping initial-condition dates, perhaps for a date chosen for a land-falling hurricane that was simultaneously chosen for the heaviest precipitation in that basin. We eliminated the overlaps, freeing up a small number for new cases to make the grand total of 1836 cases. We chose to use these cases to fill in the largest gaps in dates. Suppose the longest gap between cases was 20 days; then the first infill initial condition case date is the one in the middle of that gap. We then proceed to the next-longest gap, filling that, and so forth. The final list of cases is shown in Fig. 5. The distribution is not random. Some periods, such as the last half of 2004 were exceptionally stormy and had more than the average number of case dates. In comparison, the first half of 2023 had far fewer cases. This strategy, we hope, will allow us to train deep learning models focused on precipitation and achieve acceptable accuracy with a smaller number of cases than might be necessary with a regular sampling strategy. Such a hypothesis is difficult to test, for this would require reforecasts for both sampling strategies.

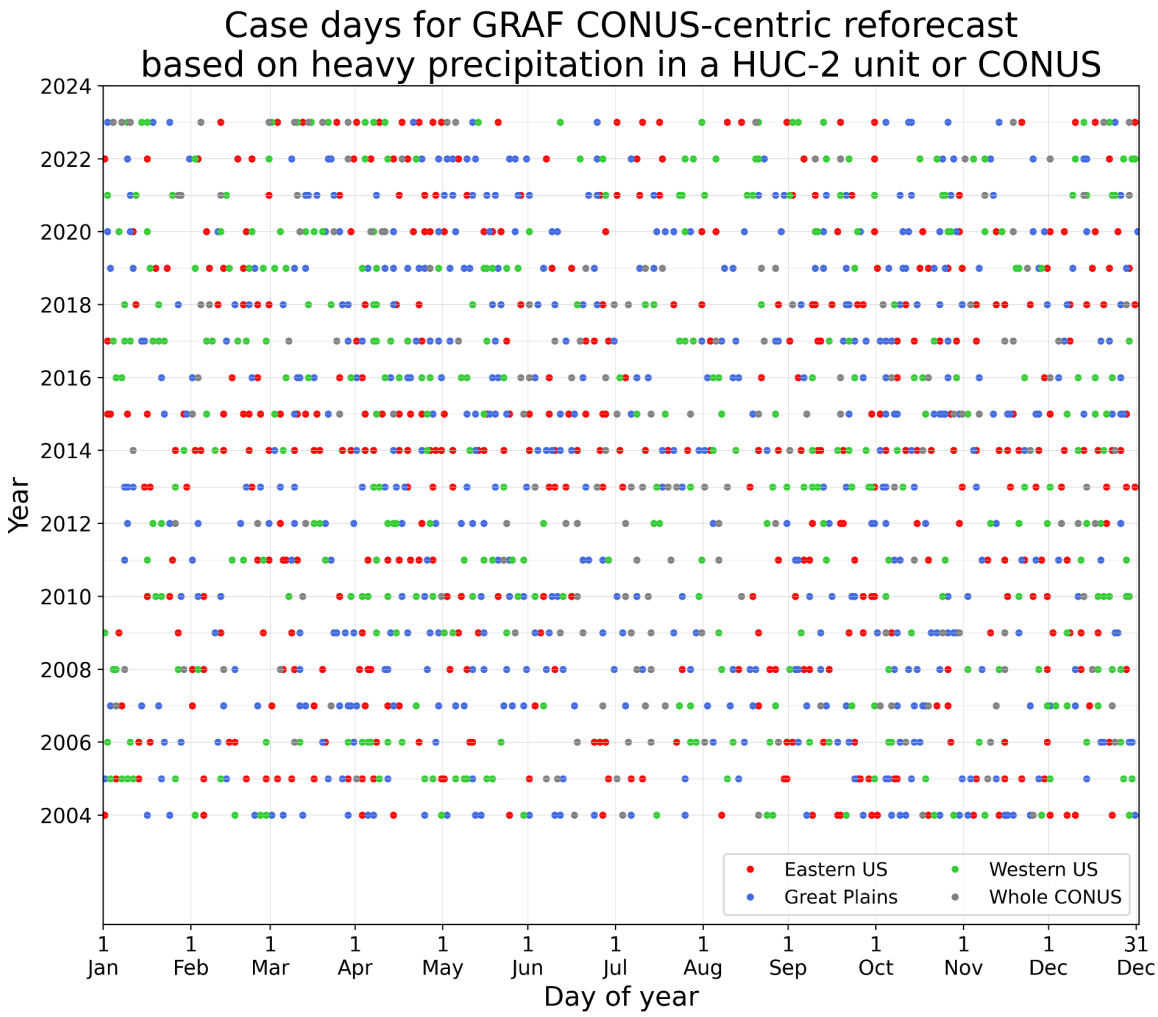

Figure 4: Case dates selected based solely on the GEFS 12-36h precipitation. These are shown for three different regions, eastern US (HUC-2 units 1-6), Great Plains (units 7-13), and western US (units 14-18), in addition to the whole CONUS.

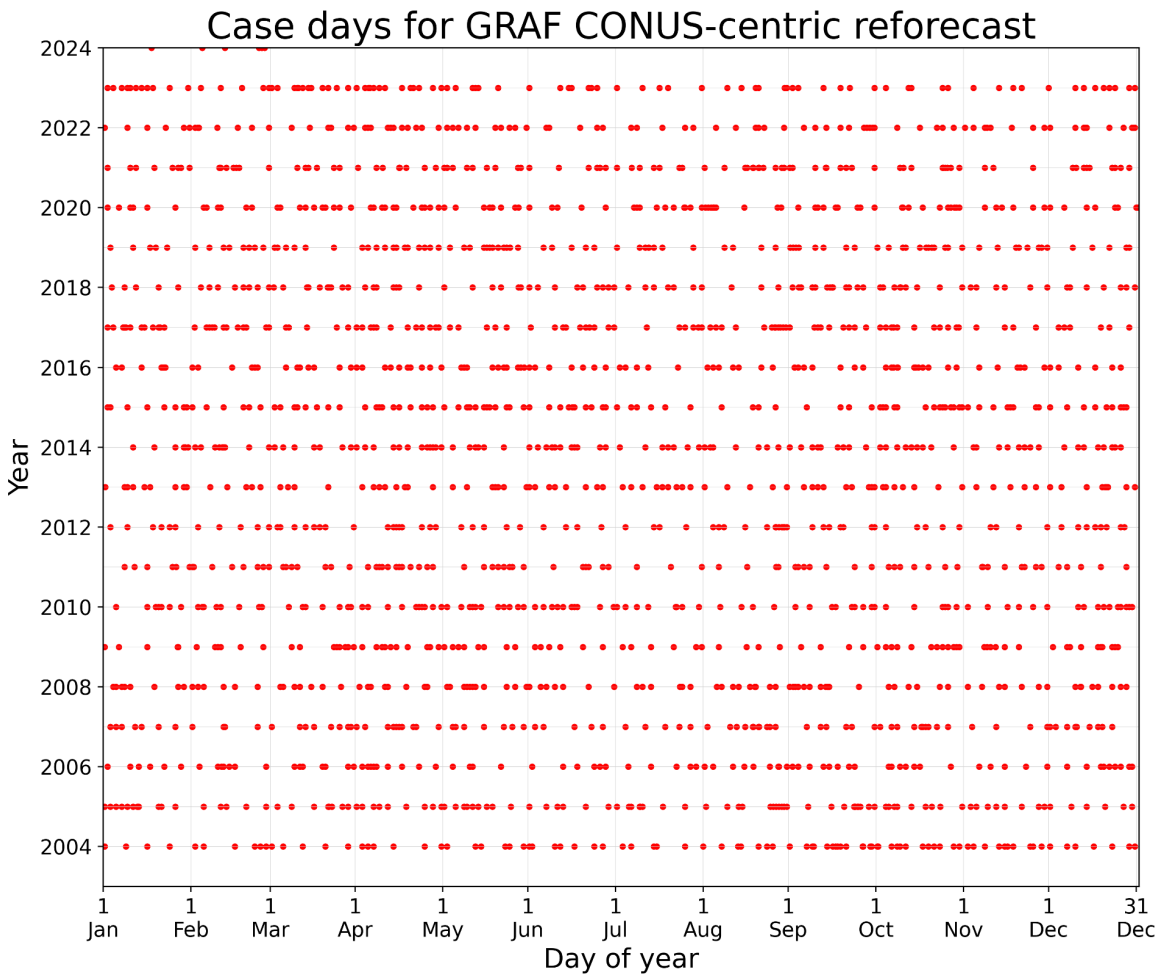

Figure 5. The final proposed list of GRAF reforecast case dates, including cases selected for major storms of varying types, heavy precipitation, and gap filling.

4. **Stored data from the reforecast.**

Table 3 shows the variables that were saved, their temporal resolution, and whether they were 2-D fields or 3-D. These data were stored in zarr format at s3://twc-graf-reforecast/ . The data are not spatially chunked. There are separate zarr files for the variables archived with 5-minute output data and 15-min output data.

A demonstration python script (demo_read_zarr_s3.py) for reading single-level model data and producing plots is included in the documentation. An example plot of model output from this script is shown in Fig. 6.

Please note these data are provided on the native model vertical coordinate. The user will need to perform for themselves a remapping to other coordinates such as a vertical pressure coordinate.

There are missing times in the data set; in some instances, for example, a 3-hour forecast may be unavailable while 2- and 4-hour forecasts were available. We apologize for these missing data.

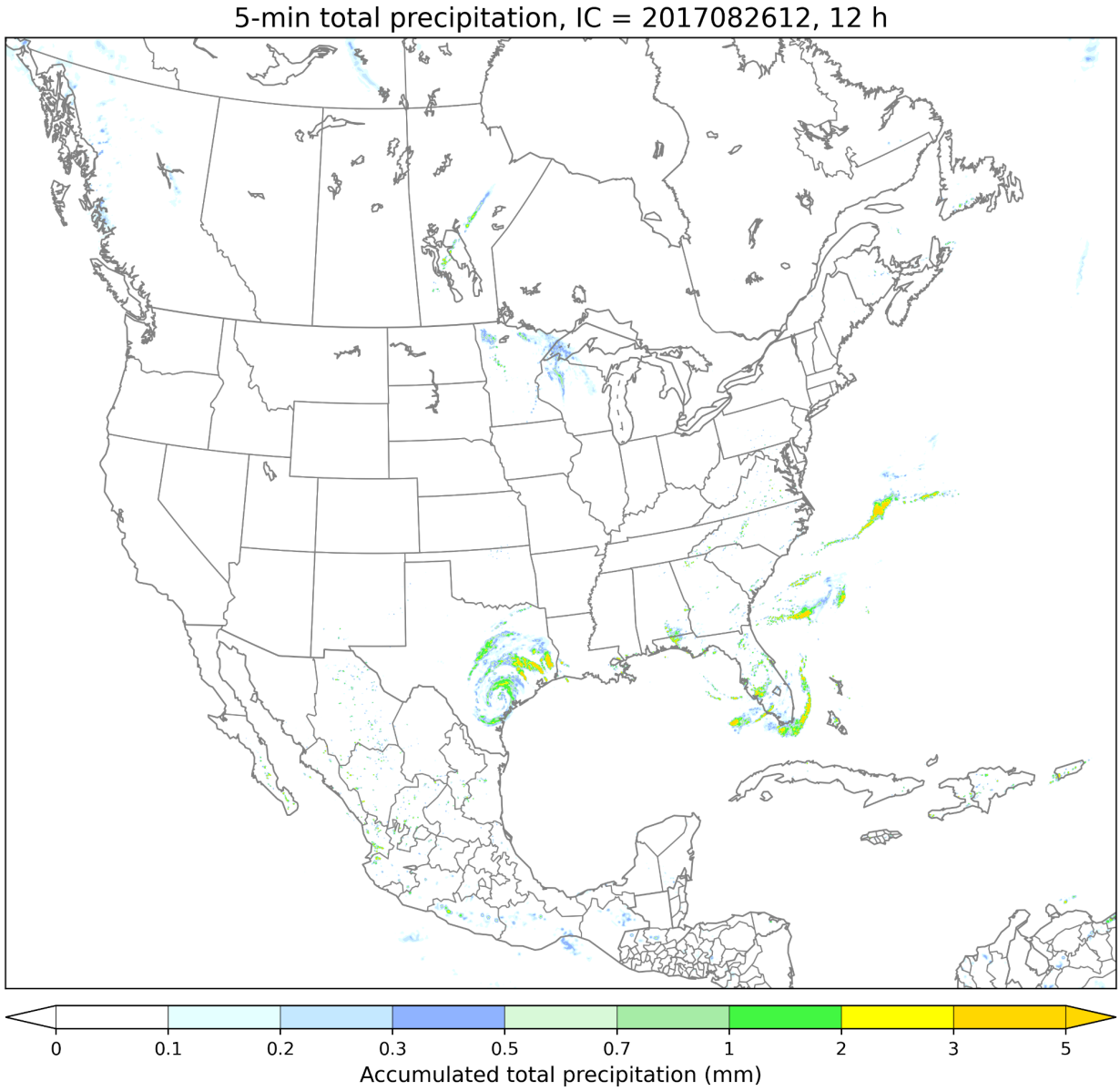

Figure 6: An illustration of 5-min precipitation amount shortly after the landfall of hurricane Harvey, produced by the python script *demo_read_zarr_s3.py*.

Acknowledgments

Some introductory text was borrowed from the NOAA Science Advisory Board, Environmental Information System Working Group's statement on deep learning numerical weather prediction (NOAA Environmental Information Systems Working Group, 2024, available here). The lead author of this document was the lead author of that statement.

References.

Andrychowicz, M., Espeholt, L., Li, D., Merchant, S., Merose, A., Zyda, F., Agrawal, S., and Kalchbrenner, N., 2023: Deep learning for day forecasts from sparse observations. *ArXiv*, https://arxiv.org/abs/2306.06079

Bauer, P., Thorpe, A. & Brunet, G., 2015: The quiet revolution of numerical weather prediction. *Nature,* **525**, 47–55. https://doi.org/10.1038/nature14956 .

Bodnar, C., W. P. Bruinsma, A. Lucic, M. Stanley, J. Brandstetter, P. Garvan, M. Riechert, J. Weyn, H. Dong, A. Vaughan, J. K. Gupta, K.Tambiratnam, A. Archibald, E. Heider, M. Welling, R. E. Turner, P. Perdikari, 2024: Aurora: a foundation model of the atmosphere. *ArXiv*, https://arxiv.org/abs/2405.13063 .

Bostrom, A., and others, 2024: Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences. *Risk Analysis*, 44, 1498–1513 DOI: 10.1111/risa.14245.

Chen, L., X. Zhong, F. Zhang, *et al.*, 2023: FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim. Atmos. Sci.*, **6**, 190. https://doi.org/10.1038/s41612-023-00512-1

Clough, S.A., M.W. Shephard, E.J. Mlawer, J.S. Delamere, M.J. Iacono, K. Cady-Pereira, S. Boukabara, P.D. Brown, 2005: Atmospheric radiative transfer modeling: a summary of the AER codes, *J. Quant. Spectrosc. Radiat. Transfer*., **91**, 233-244.

Dueben, P. D., and P. Bauer, 2018: Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11(10), 3999–4009. DOI: https://doi.org/10.5194/gmd-11-3999-2018 .

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model, *J. Geophys. Res.*, **108**, 8851, doi:10.1029/2002JD003296, D22.

Fowler, L.D., M.C. Barth, and K. Alapaty, 2020: Impact of scale-aware deep convection on the cloud liquid and ice water paths and precipitation using the Model for Prediction Across Scales (MPASv-5.2). *Geosci. Model Dev*., **13**, 2851-2877, https://doi.org/10.5194/gmd-13-2851-2020.

Galarneau, T. J., L. F. Bosart, and R. S. Schumacher, 2010: Predecessor rain events ahead of tropical cyclones. *Mon. Wea. Rev.*, **138**, 3272–3297, https://doi.org/10.1175/2010MWR3243.1.

Guan, H., and Coauthors, 2022: GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Mon. Wea. Rev.*, **150**, 647–665, https://doi.org/10.1175/MWR-D-21-0245.1.

Hamill, T. M., and others, 2021: The reanalysis for the Global Ensemble Forecast System, version 12. *Mon. Wea. Rev.*, **150**, 59-79.

Han, T., S. Guo, F. Ling, K. Chen, J. Gong, J. Luo, J. Gu, K. Dai, W. Ouyang, L. Bai, 2024: FengWu-GHR: Learning the kilometer-scale medium-range global weather forecasting. *ArXiv*, https://arxiv.org/abs/2402.00059.

Heinzeller, D., M.G. Duda, and H. Kunstmann, 2016: Towards convection-resolving, global atmospheric simulations with the Model for Prediction Across Scales (MPAS) v3.1: an extreme scaling experiment. *Geosci. Model Dev*., **9**, 77-110, 2016, doi:10.5194/gmd-9-77-2016.

Hersbach H., Bell B, Berrisford P, et al., 2020: The ERA5 global reanalysis. *Quart J Royal Meteor. Soc*., **146**, 1999–2049. https://doi.org/10.1002/qj.3803 .

Hong, S.-Y. and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Met. Soc.*, **42** (2) 2006, 129-151.

Kochkov, D., J. Yuval, I. Langmore, P. Norgaard, J. Smith, G. Mooers, M. Klower, J. Lottes, S. Rasp, P. Duben, S. Hatfield, P. Battaglia, A. Sanchez-Gonzalez, M. Willson, M. P. Brenner, and S. Hoyer, 2024: Neural general circulation models for weather and climate. *ArXiv*, https://arxiv.org/abs/2311.07222.

Lang, S., M. Alexe, M. Chantry, J. Dramsch, F. Pinault, B. Raoult, Z. Ben Bouallègue, M. Clare, C. Lessig, L. Magnusson, A.P. Nemesio, 2023: AIFS: a new ECMWF forecasting system. *ECMWF Newletter,* **178**, Winter 2023-2024. doi: 10.21957/1a8466ec2f .

Mardani, M., N. Brenowitz, Y. Cohen, J. Pathak, C.-Y. Chen, C.-C. Liu, A. Vahdat, K. Kashinath, J. Kautz, M. Pritchard, 2023: Generative residual diffusion modeling for km-scale atmospheric downscaling. *ArXiv*, https://arxiv.org/abs/2309.15214v2 .

NOAA Environmental Information Systems Working Group, 2024: *Statement on NOAA Investment in Deep Learning Numerical Weather Prediction*. NOAA Science Advisory Board, 6 pp. Available at https://sab.noaa.gov/wp-content/uploads/EISWG-Statement-on-Deep-Learning-NWP_Final_06-18-2024.pdf

Lam, R., A. Sanchez-Gonzalez, and others, 2023: GraphCast: Learning skillful medium-range global weather forecasting. *ArXiv*, https://arxiv.org/abs/2212.12794.

Li, L., R. Carver, I. Lopez-Gomez, F. Sha, J. Anderson, 2023: SEEDS: emulation of weather forecast ensembles with diffusion models. *ArXiv*, https://arxiv.org/abs/2306.14066 .

Mlawer, E.J., S.J. Taubman, P.D. Brown, M.J. Iacono and S.A. Clough, 1997: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res*., **102**, 16,663-16,682.

Nguyen, T., J. Brandstetter, A. Kapoor, J. K. Gupta, A. Grover, 2023: ClimaX: A foundation model for weather and climate. *ArXiv*, https://arxiv.org/abs/2301.10343

Park, S.-H., J. B. Klemp and W. C. Skamarock, 2014: A comparison of mesh refinement in the global MPAS-A and WRF models using an idealized normal-mode baroclinic wave simulation. *Mon. Wea. Rev*., **142**, 3614-3634. doi:10.1175/MWR-D-14-00004.1

Pathak, J., S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, P. Hassanzadeh, K. Kashinath, and A. Anandkumar, 2022: FourCastNet: A global data-driven high-resolution weather model using adaptive fourier neural operators. ArXiv, https://arxiv.org/abs/2202.11214 .

Price, I., A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, R. Lam, M. Willson, 2024: GenCast: Diffusion-based ensemble forecasting for medium-range weather. *ArXiv*, https://arxiv.org/abs/2312.15796.

Rasmussen, R. M., and Coauthors, 2023: CONUS404: The NCAR–USGS 4-km long-term regional hydroclimate reanalysis over the CONUS. *Bull. Amer. Meteor. Soc.*, **104**, E1382–E1408, https://doi.org/10.1175/BAMS-D-21-0326.1.

Sandbach, S., J. Thuburn, D. Vassilev, and M. G. Duda, 2015: A semi-implicit version of the MPAS-atmosphere dynamical core. *Mon. Wea. Rev*., **143**, 3838–3855. doi:10.1175/MWR-D-15-0059.1

Seaber, P.R., Kapinos, F.P., and Knapp, G.L., 1987, Hydrologic Unit Maps: U.S. Geological Survey Water-Supply Paper 2294, 63 p.

Skamarock, W. C., J. B. Klemp, M. G. Duda, L. D. Fowler, S. Park, and T. D. Ringler, 2012: A Multiscale Nonhydrostatic Atmospheric Model Using Centroidal Voronoi Tesselations and C-Grid Staggering. *Mon. Wea. Rev.*, **140**, 3090–3105, https://doi.org/10.1175/MWR-D-11-00215.1.

Vannitsem, S., D. S. Wilks and J. W. Messner, 2019: *Statistical Postprocessing of Ensemble Forecasts*. Elsevier Press, DOI: https://doi.org/10.1016/C2016-0-03244-8, 347 pp.

Wang, W., 2022: Forecasting convection with a “scale-aware” Tiedtke cumulus parameterization scheme at kilometer scales. *Wea. Forecasting*, **37**, 1491–1507, https://doi.org/10.1175/WAF-D-21-0179.1.

Weyn, J.A., D. R. Durran, and R. Caruana, 2020: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *J. Adv. in Modeling Earth Sys*., **12**, DOI: https://doi.org/10.1029/2020MS002109 .

Weyn, J.A., D. R. Durran, R. Caruana, and N. Cresswell-Clay, 2021: Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *J. Adv. in Modeling Earth Sys.*, **13**, e2021MS002502. DOI: 10.1029/2021MS002502 .

Zhu, Y., and R. E. Newell, 1994: Atmospheric rivers and bombs. *Geophys. Res. Letters*, **21** (18), 1999–2002. doi:10.1029/94GL01710.

Variable	Variable name(s) in zarr files	Temporal frequency	2D or 3D	Units
Water vapor mixing ratio	qv	15 min	3D (atm)	kg/kg
Cloud water mixing ratio	qc	15 min	3D (atm)	kg/kg
Rain water mixing ratio	qr	15 min	3D (atm)	kg/kg
Ice mixing ratio	qi	15 min	3D (atm)	kg/kg
Snow mixing ratio	qs	15 min	3D (atm)	kg/kg
Graupel mixing ratio	qg	15 min	3D (atm)	kg/kg
u-wind component	uReconstructZonal	15 min	3D (atm)	m/s
v-wind component	ureconstructMeridional	15 min	3D (atm)	m/s
w-wind component	w	15 min	3D (atm)	m/s
Total pressure	p	15 min	3D (atm)	Pa
Temperature	temperature	15 min	3D (atm)	K
Potential temperature	theta	15 min	3D (atm)	K
Soil temperature	tsl	15 min	3D (soil)	K
Soil equivalent liquid water	sh2o	15 min	3D (soil)	m^3/m^3
Soil moisture	smois	15 min	3D (soil)	m^3/m^3
Kuchera snow ratio	snow_ratio	15 min	2D	m/m
Total snow depth	snowh	15 min	2D	m
Visibility	visibility	15 min	2D	m
Conditional probability of rain	cporain	15 min	2D	n/a
Conditional probability of snow	cposnow	15 min	2D	n/a
Conditional probability of ice	cpoice	15 min	2D	n/a
10-m u component	u10	15 min	2D	m/s
10-m v component	v10	15 min	2D	m/s
All-sky downward surface shortwave radiation flux	swdnb,	15 min	2D	W/m^2
Downward surface shortwave Direct Normal Flux	swdnbdn	15 min	2D	W/m^2
Downward all-sky surface flux, short and longwave, 1-h average	swdnb01h	15 min	2D	W/m^2
Downward surface shortwave Direct Normal Flux, 1-h average	swdnbdn01h	15 min	2D	W/m^2
Precipitation rate	prate	5 min	2D	mm/h
Predominant precipitation type	ptype	5 min	2D	n/a
rain accumulation	rain_bucket	5 min	2D	m

convective rain accumulation	convective_bucket	5 min	2D	m
ice accumulation	zrain_bucket	5 min	2D	m
snow accumulation	snow_bucket	5 min	2D	m
total precipitation accumulation	apcp_bucket	5 min	2D	m
Total cloud cover	total_cloud_cover	5 min	2D	%
Mean sea-level pressure	mslp	15 min	2D	Pa
2-m temperature	t2m	15 min	2D	K
2-m dewpoint	dewpoint_2m	15 min	2D	K
2-m specific humidity	q2	15 min	2D	kg/kg
Total-column precipitable water	precipw	15 min	2D	kg/m^2
Skin temperature, including SST	skintemp	15 min	2D	K
Wind gust	windgust10m	15 min	2D	m/s
All-sky top of atmosphere outgoing longwave	olrtoa	15 min	2D	W/m^2
Lifted index	bli	15 min	2D	K
Convective available potential energy	cape	15 min	2D	J/kg
Convective inhibition	cin	15 min	2D	J/kg
Lifted condensation level	lcl	15 min	2D	m
Ceiling above ground level	ceiling_agl	15 min	2D	m
Echo top (18 dBz)	echotop	15 min	2D	m
Fire weather index	fwi	15 min	2D	n/a
PBL height	hpbl	15 min	2D	m
Latent heat at the surface	lh	15 min	2D	W/m^2
Hourly averaged latent heat flux	lh01h	15 min	2D	W/m^2
Power disruption index	pdi	15 min	2D	n/a
Thunderstorm potential index	tpi	15 min	2D	n/a

Table 3. List of forecast output variables for the GRAF reforecast, their temporal frequency, and whether the data are 2-D or 3-D fields. “Atm” indicates for atmospheric levels, “soil” for soil levels.