WeatherBench 2: A benchmark for the next generation of data-driven global weather models

Stephan Rasp[1,*], Stephan Hoyer[1], Alexander Merose[1], Ian Langmore[1], Peter Battaglia[2], Tyler Russell[1], Alvaro Sanchez-Gonzalez[2], Vivian Yang[1], Rob Carver[1], Shreya Agrawal[1], Matthew Chantry[3], Zied Ben Bouallegue[3], Peter Dueben[3], Carla Bromberg[1], Jared Sisk[1], Luke Barrington[1], Aaron Bell[1], and Fei Sha[1]

[1]Google Research

[2]Google DeepMind

[3]European Centre for Medium-Range Weather Forecasts

[*]Corresponding author: srasp@google.com

Abstract

WeatherBench 2 is an update to the global, medium-range (1–14 day) weather forecasting benchmark proposed by Rasp et al. (2020), designed with the aim to accelerate progress in data-driven weather modeling. WeatherBench 2 consists of an open-source evaluation framework, publicly available training, ground truth and baseline data as well as a continuously updated website with the latest metrics and state-of-the-art models: `https://sites.research.google/weatherbench`. This paper describes the design principles of the evaluation framework and presents results for current state-of-the-art physical and data-driven weather models. The metrics are based on established practices for evaluating weather forecasts at leading operational weather centers. We define a set of headline scores to provide an overview of model performance. In addition, we also discuss caveats in the current evaluation setup and challenges for the future of data-driven weather forecasting.

1 Introduction

Global, medium-range (1–14 day) numerical weather prediction (NWP) is a key component of modern weather forecasting (Bauer et al., 2015). It has huge economic and societal impact as many significant weather events occur on this time scale, such as heat waves, tropical- and extra-tropical cyclones, droughts or heavy precipitation leading to flooding. In addition to providing valuable forecasts, global NWP models serve a range of additional purposes: they provide boundary conditions for regional, high-resolution models; they are used to create (re-)analyses; and they serve researchers as tools to better understand the atmosphere. Current NWP models are based on discretizations of the governing equations describing fluid flow and thermodynamics (Kalnay, 2002). As of 2023, most global models have a horizontal grid spacing of less than 25 km. The European Center for

Medium-range Weather Forecasting's (ECMWF) Integrated Forecast System (IFS) model, for example, has a resolution of 0.1°, around 9km, in its high-resolution (HRES) and, since the 2023 upgrade, also in its ensemble (ENS) configuration. This still leaves many important physical processes unresolved, for example cloud physics and radiation. The effect of these processes on the resolved scales has to be approximated in so-called parameterizations (Stensrud, 2007). Another important aspect of NWP is estimating the current state of the atmosphere, required to initialize model forecasts. This is done using data assimilation algorithms that combine model forecasts and observations to produce an analysis, a best guess of the current state of the atmosphere. The last few decades have seen a steady improvement in global NWP driven by increased computing power, which in turn allowed higher resolution and more ensemble members, better observations and data assimilation, and better representations of the physical processes (Magnusson and Källén, 2013). Despite the impressive progress in physical global NWP, however, there is still ample room for improvement. Recent studies estimate that the intrinsic limit of predictability of mid-latitude weather is at around 15 days, while the current practical limit of predictability is at around 10 days. Around half of the remaining 5 days of potential skill stem from model improvements, the other half from better initial conditions (Zhang et al., 2019; Selz et al., 2022).

In the last few years, spurred by the rise of artificial intelligence (AI) in domains like computer vision and natural language processing (LeCun et al., 2015), researchers have been exploring the possibility of using modern machine learning algorithms for weather forecasting. A number of initial attempts at building data-driven medium-range NWP models (Dueben and Bauer, 2018; Scher, 2018; Weyn et al., 2019) led to the creation of WeatherBench (Rasp et al., 2020, from here on called WB1), a benchmark for global, medium-range weather prediction. The goal of WB1 was to provide a common, reproducible framework for evaluating global, data-driven forecasts and compare them against traditional baselines. Benchmarks have been hugely influential in the ML community to measure progress on specific tasks. Famous examples include ImageNet (Deng et al., 2009) which helped kick-start the AI revolution in computer vision (Krizhevsky et al., 2017) and the GLUE benchmark (Wang et al., 2018) in language. After its release, WB1 was used by a number of studies to explore different machine learning approaches. Weyn et al. (2020) used a cubed sphere projection in combination with a UNet architecture to iteratively predict the atmospheric state at 2° resolution. Rasp and Thuerey (2021) used a deep Resnet architecture to directly predict fields up to 5 days ahead at a 5.625° resolution. Clare et al. (2021) similarly used a Resnet but added a probabilistic output layer. A probabilistic extension to WeatherBench along with several ML baselines was published as well (Garg et al., 2022). An up-to-date leaderboard of WB1 metrics can be found at `https://github.com/pangeo-data/WeatherBench`. These studies, all using comparatively coarse resolution, did not come close to the skill of current physical NWP models.

In early 2022, improvements in data-driven weather models considerably picked up in pace. Keisler (2022) used a graph neural network (GNN; Pfaff et al., 2021) iteratively with a time step of 6 hours at a resolution of 1° and 13 vertical levels. On deterministic upper-level verification metrics the model achieves skill comparable to some operational NWP models. Pathak et al. (2022) used a modified vision transformer (Guibas et al., 2022), called FourCastNet, for prediction at a very high resolution of 0.25° and a 6 hour time step. Pangu-Weather (Bi et al., 2023), based on a different variation of vision transformers at equally high-resolution, obtained deterministic metrics that outperform HRES. Shortly after, GraphCast (Lam et al., 2022) built upon the work of Keisler (2022) and scaled a GNN to 0.25° horizontal resolution. On many deterministic metrics as well as extreme weather indicators, GraphCast also outperforms HRES. In 2023, another vision transformer variant called FengWu (Chen et al., 2023a) was published with state-of-the-art, deterministic scores for longer forecast lead times. Similarly, the FuXi model (Chen et al., 2023c) achieves deterministic scores similar to the IFS ensemble mean up to 15 days. SwinRDM (Chen et al., 2023b) combines a recurrent network with a diffusion model to provide granular predictions at 0.25° resolution.

In light of the rapid advances of the field, the need for an updated benchmark that allows for easy comparison between different approaches becomes apparent. This paper will lay out the design principles for WB2, followed by a detailed description of the evaluation metrics and datasets, as well as results on the headline scores. We will also discuss several issues with WB2 and potential future directions for this benchmark dataset.

2 Design decisions for WeatherBench 2

A general challenge of designing benchmarks for weather prediction is that weather is a very high-dimensional and multi-faceted problem (Dueben et al., 2022). Every use case has slightly different requirements and quality measures. Take as an example model development at ECMWF, one of the world's leading operational NWP centers. ECMWF tracks a large number of metrics to evaluate their model performance. The headline scores[1] and scorecard[2] serve as concise summaries but are only the tip of the iceberg. In addition to quantitative scores, feedback is also be collected from in-house experts and end users, each of them interested in specific aspects of the forecast system. Recognizing the limitations of particular scores is even more important to keep in mind when evaluating ML-based approaches, which often violate standard assumptions in traditional NWP. All this is to say that no single metric or set of metrics will ever be able to fully describe what a "good" forecast is. WB2 does not attempt to do so. Similarly to ECMWF, WB2 defines a set of headline scores (described in Section 5) and evaluation tools that aim to capture key aspects of medium-range weather forecasting but are not meant to be exhaustive. Therefore, WB2 should not be seen as a traditional benchmark challenge with a single leaderboard but rather as a tool to compare different approaches on different aspects.

While WB2 defines the targets to be evaluated, it leaves the remaining modeling setup open (e.g., which inputs or model resolution to use). In this sense, WB2 is a benchmark for entire forecast systems, i.e., the combination of choice of input data, training setup and model architecture, rather than focusing on the models specifically. While this makes it harder to compare different model architectures, we believe that designing the non-model part of the forecast system components is equally as important and different approaches should be encouraged to accelerate progress in the field, as long as overfitting is avoided. However, it is important to recognize that different design choices can lead to advantages and disadvantages. One important choice is which input dataset to use for starting ("initializing") forecasts. Currently, most models use re-analysis datasets, in particular ERA5, which would not be available in an operational setup and provides an potential advantage over operational initial conditions (more details in Section 6.2).

Another priority of WB2 is to emphasize the importance of probabilistic prediction. Weather forecasting is an inherently uncertain endeavor because of chaotic error growth (Lorenz, 1963; Zhang et al., 2007), implying that even with perfect models and near perfect initial conditions there is a range of possible outcomes. In operational NWP, this fact led to the development of ensemble prediction systems in the 1990s (Palmer et al., 1993; Toth and Kalnay, 1993). In an ensemble, a number of forecasts (typically 10–100) are run with perturbed initial conditions and sometimes model physics, providing different potential realizations of future weather. One of the most important advantages of ensemble forecasts is that they can provide reliable information for decision making. For example, they can be used to estimate the probability of extreme events that a single,

[1] https://www.ecmwf.int/en/forecasts/quality-our-forecasts

[2] https://sites.ecmwf.int/ifs/scorecards/

Model/ Dataset	**Type**	**Initial conditions**	Δx	**Levels**	**Training data**	**Training resources**	**Inference time**
ERA5	Reanalysis		0.25°	137			
IFS HRES	Forecast	Operational	0.1°	137			∼ 50 minutes (*)
IFS ENS	Forecast	Operational	0.2°	137			
ERA5 forecasts	Hindcast	ERA5	0.25°	137			
Keisler (2022)	Forecast	ERA5	1°	13	ERA5 (35 years, see text)	5.5 days; one A100 GPU	∼1 second; single GPU
Pangu-Weather	Forecast	ERA5	0.25°	13	ERA5 (1979-2017)	16 days; 192 V100 GPUs	several seconds; single GPU
GraphCast	Forecast	ERA5	0.25°	37	ERA5 (1979-2019)	4 weeks; 32 TPU v4	∼1 minute; single TPU

Table 1: Table of datasets and models; Δx refers to the horizontal resolution and "levels" to the number of vertical model levels used in the input and output. (*) See details in the text for IFS inference time.

deterministic forecast might miss. Tropical cyclone track forecasts (often displayed as plumes) are a prime example of this. The data-driven models discussed above, with the exception of a perturbed ensemble experiment in the Pangu-Weather paper, all only provide deterministic forecasts. Similarly to the evolution of dynamical forecast systems, data-driven forecasts will need to become probabilistic to provide users with the most actionable information. This could be possible as a post-processing step on top of deterministic models, for example, through ensemble dressing or lagged forecasts. However, the success of NWP ensembles hints at the advantages of designing probabilistic forecast systems from the ground up, especially for capturing spatio-temporal correlations. For this reason, WB2 will include operational probabilistic verification metrics and baselines from the start. In data-driven modeling, there are several ways to create probabilistic forecasts, from predicting parameterized marginal distributions directly (see e.g. Andrychowicz et al., 2023) to producing generative roll-outs. Some aspects of how to evaluate these forecasts are discussed in Section 6.3.

Finally, WB2 will provide a dynamic, open-source framework that can evolve with the needs of the ML-weather community. All data and code needed for evaluation are publicly available and can be extended to accommodate more detailed evaluation in the future, driven either by us or community contributors.

3 Data, baselines and data-driven models

Most of the datasets below are available in on Google Cloud Storage in Zarr format. For the latest details on these datasets, please visit `https://weatherbench2.readthedocs.io/en/latest/data-guide.html`. Table 1 provides a summary of key facts for each of the datasets used.

3.1 ERA5

The ERA5 dataset (Hersbach et al., 2020) is used as the ground truth dataset for WB2 and as the training dataset for many of the data-driven approaches described above. ERA5 is a reanalysis dataset based on a 0.25° (roughly 30 km) version of ECWMF's HRES model operational in 2016 (cycle 42r1) and ECMWF's 4D-Var data assimilation which uses a wide range of direct and remote sensing observations. ERA5 uses 12 hour assimilation windows from 21–09 and 09-21 UTC, during which a "prior" forecast initialized from the previous assimilation window is combined with observations to produce an analysis, a "best guess" of the Earth system state during this window. ERA5 data is available at hourly resolution from 1940 to present at the Copernicus Climate Data Store[3]. A subset of the ERA5 dataset is available on the cloud in the cloud-optimized Zarr format.

Note that using ERA5 for evaluation and training has some caveats. First, while attempting to be close to observations, ERA5 is a model simulation which is closer to the truth for some variables than for others. Especially for precipitation, ERA5 sometimes shows large differences to rain gauge measurements Lavers et al. (2022). We still include precipitation evaluation based on ERA5 here but advise caution when interpreting those results (for more details, see discussion in Section 6.1). Second, ERA5 uses a longer assimilation window compared to operational forecasts (see below). This helps to better constrain the guess of the atmospheric state but would delay the initialization of forecasts in a real time setting. At 00 UTC, for example, ERA5's assimilation window goes 9 hours "into the future". Operationally, the data assimilation window only extend 3 hours ahead from the forecast initialization time. One could use 06/18UTC initialization to level the playing field a little bit more. This issue has been extensively discussed in the supplement of (Lam et al., 2022, Fig. 10) who found that forecasts initialized from ERA5 at 00/12 UTC indeed perform better than those initialized at 06/18UTC, when the ERA5 analysis also only has a 3 hour look-ahead. The difference in RMSE for GraphCast was up to 5%. However, the 06/18UTC operational analysis are also produced using fewer DA cycles compared to the 00/12UTC analysis. For WB2, we chose to evaluate forecasts initialized at 00/12UTC. The main reasons for this are that some key baselines (IFS ENS and the ERA5 forecasts) were only available for 00/12 UTC initializations and that this has been the standard for most other ML-based evaluations so far, which allows us to directly include some of them here.

3.2 Climatology

The climatology is used for computing certain skill scores, in particular ACC and SEEPS, and as a baseline forecast. Here we follow Jung and Leutbecher (2008) and compute the climatology c as a function of the day of year (doy) and the time of day (tod) by taking the mean of ERA5 data from 1990 to 2019 (inclusive) for each grid point. A sliding window of 61 days is used around each *doy-tod* combination with weights linearly decaying to zero from the center. This removes sample noise and makes the climatology smoother in time, at the expense of reducing the seasonal amplitude.

A probabilistic version of the climatology is created by taking each of the 30 years from 1990 to 2019 as an ensemble member, this time without smoothing.

Note that a 30 year climatology will include some climate drift, especially for temperature. Here, we do not apply any measure to correct for this.

3.3 IFS HRES

Our main baseline comes from ECMWF's operational IFS model. ECMWF's forecasts are widely regarded as one of the best medium-range NWP models[4]. Since 2016, the IFS in its HRES configuration has been run at 0.1° (roughly 9 km) horizontal resolution. The operational model is

[3] `https://cds.climate.copernicus.eu/`

[4] `https://wmolcdnv.ecmwf.int/`

updated regularly, approximately once to twice a year, which means that the exact model configuration might change during the evaluation period. Usually, updates are associated with slight improvements in most evaluation metrics, though not all. However, changes in the IFS are typically gradual. A comprehensive model description can be found at `https://www.ecmwf.int/en/publications/ifs-documentation`. A schedule of model upgrades can be found at `https://confluence.ecmwf.int/display/FCST/Changes+to+the+forecasting+system`. Initial conditions are created every 6 hours using an ensemble 4D-Var system using information from the previous assimilation cycle's forecast as well as observations in a +/- 3 hour window. After accounting for the time to perform data assimilation and forward simulation, forecasts have a latency of 5.75 to 7 hours from the time at which they are initialized[5]. Forecasts started at 00 and 12 UTC are run up to a lead time of 10 days. 06 and 18 UTC initializations are run for 3.75 days.

Running a 15 day TCO1279 ($\sim$ 9km) simulation in the setup used for the operational ensemble (using the new high-resolution setup) takes around 52 minutes with I/O and 46 minutes without I/O on 64 128-core (AMD EPYC Rome) nodes.

3.4 IFS HRES Initial Conditions

For evaluating IFS forecasts, we use operational analyses as ground truth rather than ERA5. This is because evaluating IFS forecasts against ERA5 would result in a non-zero error at time step $t = 0$ and therefore would put the IFS at an unfair disadvantage compared to data-driven models trained on and evaluated against ERA5. The difference is most pronounced during the early lead times and becomes negligible at longer lead times. See Fig. S6 for a comparison.

Here, as in Lam et al. (2022) we use the initial conditions, i.e. the IFS HRES forecasts at t=0 as our "analysis". Note that this dataset is slightly different from the official analysis product on the ECMWF archive. This is because for the official analysis product an additional surface data assimilation step is taken. For ECMWF's internal evaluation, the official analysis product is used. However, the qualitative differences in the evaluation scores resulting from this discrepancy should be small except for the first time steps after initializations and are mostly limited to surface temperature. For very short lead times, results should not be over-interpreted anyway. Note that for precipitation accumulations we use ERA5 as a precipitation ground truth for all models.

3.5 IFS ENS

ECMWF also runs an ensemble version (ENS) of IFS, until to the 2023 upgrade at 0.2° resolution (including the 2020 evaluation period used here), now at 0.1° resolution. The ensemble consists of a control run and 50 perturbed members. The initial conditions of the perturbed members are created by running an ensemble of data assimilations (EDA) in which observation errors, model errors and boundary condition errors are represented by perturbations[6]. The difference of each of the EDA analyses to the mean is then used as a perturbation to the HRES initial conditions. In addition, to more accurately represent forecast uncertainty, singular vector perturbations are added[7] to the initial conditions. Model uncertainties during the forecasts are represented by stochastically perturbed parameterization tendencies (Buizza et al., 1999, SPPT), in which spatially and temporally correlated perturbations are added to the model physics tendencies. Ensemble forecasts are initialized at 00 and 12 UTC and run out to 15 days.

[5] `https://confluence.ecmwf.int/display/DAC/Dissemination+schedule`

[6] `https://confluence.ecmwf.int/display/FUG/Section+5.1.1+Ensemble+of+Data+Assimilations+-+EDA`

[7] `https://confluence.ecmwf.int/display/FUG/Section+5.1.2+Singular+Vectors+-+SV`

3.6 IFS ENS Mean

We also include the IFS ENS mean as a baseline, which we computed by simply averaging over the 50 members. The ensemble mean does not represent a realistic forecast but often performs very well on deterministic metrics.

3.7 ERA5 forecasts

For research purposes, ECMWF ran a set of 10 day hindcasts initialized from ERA5 states at 00/12UTC with the same IFS model version used to create ERA5 (see above). These forecast are available in the MARS archive[8]. Here we downloaded the variables required to compute the headline scores for the year 2020 (with the exception of total precipitation which was not available). Note that data until 5 days lead time is available at 6h intervals, and 12h intervals from 5 to 10 days lead time.

The ERA5 forecasts provide a like-for-like baseline for an AI model initialized from and evaluated against ERA5. They benefit from the same, longer assimilation window compared to the operational initial conditions and are run at 0.25° resolution—similar to many modern AI methods. Because of the lower resolution and older model relative to the operational IFS HRES in 2020, one would expect the operational model to be more skillful by itself. How this balances out against the difference in initial conditions will be discussed in Section 5.

3.8 Keisler (2022) Graph Neural Network

Keisler (2022) used a graph neural network architecture (Pfaff et al., 2021) with an encoder that maps the original 1° latitude-longitude grid to an icosahedron grid, on which several rounds of message-passing computations are performed, before decoding back into latitude-longitude space. The model takes as input the atmospheric states at $t = 0$ and $t = -6h$ and predicts the state at $t = 6h$. To forecast longer time horizons, the model's outputs are fed back as inputs autoregressively. The state consists of 6 three-dimensional variables at 13 pressure levels. ERA5 data is used for training with 1991, 2004 and 2017 used for validation, 2012, 2016 and 2020 for testing and the remaining years from 1979 to 2020 for training. During training the model is trained to minimize the cumulative error of up to 12 time steps (3 days). Model training took 5.5 days on a single Nvidia A100 GPU.

3.9 Pangu-Weather

Pangu-Weather (Bi et al., 2023) is a data-driven weather model based on a transformer architecture. It predicts the state of the atmosphere at $t = t + \Delta t$ based on the current state. The state is described by 5 upper-air variables and 4 surface variables on a 0.25° horizontal grid (same as ERA5) with 13 vertical levels for the upper-air variables. The model is trained using ERA5 data from 1979 to 2017 (incl.) with 2019 for validation and 2018, 2020 and 2021 for testing. Here we evaluate forecasts for 2020. Four different model versions are trained for different prediction time steps $\Delta t = \{1h, 3h, 6h, 24h\}$. To create forecasts for an arbitrary lead time model predictions are chained together autoregressively from the four different lead time models, using the fewest number of steps. For example, to create a 31h forecast, a 24h forecast is followed by a 6h and then a 1h forecast. The maximum lead time for the data used here is 7 days. Training the model took 16 days on 192 Tesla-V100 GPUs. Creating a prediction with the trained model around 1.5s on a single GPU. Inference code for Pangu-Weather can be found at `https://github.com/198808xc/Pangu-Weather`.

[8] MARS parameters: class=ea, stream=oper, expver=11, type=fc

Symbol	Range	Description
f		Forecast
o		Ground Truth
c		Climatology
t	$1, ..., T$	Verification time
l	$1, ..., L$	Lead time
i	$1, ..., I$	Latitude index
j	$1, ..., J$	Longitude index
m	$1, ..., M$	Ensemble member index

Table 2: Notation used in the evaluation metrics.

3.10 GraphCast

GraphCast (Lam et al., 2022) is similar in structure to Keisler (2022) but operates on a higher resolution input with 6 upper-level variables on a 0.25° horizontal grid with 37 vertical levels, and additionally 5 surface variables. The model is also trained autoregressively up to a time horizon of 12 time steps (3 days). Training took around four weeks on 32 TPU v4 devices. Creating a single 10-day forecast takes less than a minute on a single TPU. Here, we evaluate a version of GraphCast that was trained on ERA5 data from 1979 to 2019 (incl.). See Suppl. 5.1 of Lam et al. (2022) for details. Code for GraphCast can be found at `https://github.com/deepmind/graphcast`.

4 Evaluation protocol and metrics

The WB2 evaluation protocol sticks closely to the forecast verification used by the WMO (World Meteorological Organization, 2019) and operational weather centers like ECMWF. Table 4 provides a list of the notation used.

4.1 Evaluation time period and initialization times

In its initial version WB2 will use the year 2020. 2020 was chosen a) because it provides a compromise between recency and leaving several more recent years for independent testing if modeling groups desire to do so; and b) because data was available for many of the AI baseline models. One year is also a robust enough sample size for most of the metrics defined below. It should be noted, however, that for metrics focused on extreme events, e.g. hurricanes or heat waves, larger sample sizes are required. As WB2 is updated, evaluation can be changed to a more recent or longer period.

Evaluation is done on all 00 and 12 UTC initialization times for 2020, i.e. from January 1 2020 00UTC to December 31 12 UTC. This means that some forecasts will extend into 2021. We compared this to evaluating only forecast lead times that are valid in 2020 but the difference in the scores was negligible. The evaluation time step can be freely chosen. We use 6 hours as our highest resolution.

4.2 Evaluation resolution and area

Before the computation of the metrics, all forecasts and ground truths are first-order conservatively regridded to 1.5° resolution.[9] 1.5° is also used as the standard resolution for evaluation by

[9] A regridding tool is available on the WB2 GitHub page.

the WMO (World Meteorological Organization, 2019) and at ECMWF. Regridding to a common lower resolution allows all models to be evaluated without penalizing coarser-resolution models. Absolute metric values can differ significantly between resolutions but the relative differences between different models are consistent across different evaluation resolution. For a detailed analysis, see Supplement Figs. S11, S12 and S13. However, this should not give the false impression that higher-resolution models are not more accurate as relevant small scale details can be resolved.

All results shown in this paper are computed over all global grid points. Additionally, scores computed many other regions are shown on the website: `https://sites.research.google/weatherbench`.

A note on below-ground grid points: For some pressure level variables (e.g. 850hPa temperature), some grid points in high-altitude regions will be "below" ground; i.e., the surface pressure is actually smaller than the pressure level. ERA5 and IFS output will still provide interpolated values at these locations, even though they do not correspond to real physical variables. Since it has been common practice to use these interpolated values in ML training and evaluation they are be included here as well.

4.3 Deterministic metrics

All metrics are computed using an area-weighting over grid points. This is because on an equiangular latitude-longitude grid, grid cells at the poles have a much smaller area compared to grid cells at the equator. Weighing all cells equally would result in an inordinate bias towards the polar regions. The latitude weights $w(i)$ are computed as:

$$w(i) = \frac{\sin\theta_i^{\mathrm{u}} - \sin\theta_i^{\mathrm{l}}}{\frac{1}{I}\sum_i^I(\sin\theta_i^{\mathrm{u}} - \sin\theta_i^{\mathrm{l}})}, \tag{1}$$

where θ_i^u and θ_i^l indicate upper and lower latitude bounds, respectively, for the grid cell with latitude index i. All vertical (pressure) levels are treated separately. For readability, no level index is included in the equations below.

4.3.1 Root mean squared error (RMSE)

The RMSE is defined for each variable and level as

$$\mathrm{RMSE}_l = \frac{1}{T}\sum_t^T \sqrt{\frac{1}{IJ}\sum_i^I\sum_j^J w(i)(f_{t,l,i,j} - o_{t,i,j})^2} \tag{2}$$

Note that in our definition of the RMSE the time mean is outside of the square root. Many other definitions, e.g. that used by the WMO, include the time mean inside of the root. We compared both versions and found differences to be small (less than 2% absolute difference on average).

The wind vector (WV) RMSE is computed as

$$\mathrm{RMSE}_l^{\mathrm{WV}} = \frac{1}{T}\sum_t^T \sqrt{\frac{1}{IJ}\sum_i^I\sum_j^J w(i)\left[(u_{t,l,i,j}^f - u_{t,i,j}^o)^2 + (v_{t,l,i,j}^f - v_{t,i,j}^o)^2\right]} \tag{3}$$

where u^f, u^o, v^f and v^o are the u and v components of wind in the forecast and observations, respectively.

4.3.2 Anomaly correlation coefficient (ACC)

The ACC is computed as the Pearson correlation coefficient of the anomalies with respect to the climatology c:

$$f'_{t,l,i,j} = f_{t,l,i,j} - c_{t,l,i,j}; \quad o'_{t,i,j} = o_{t,i,j} - c_{t,i,j} \tag{4}$$

where $c_{t,l,i,j}$ and $c_{t,i,j}$ indicate the climatology corresponding to the appropriate *tod-doy* combination. The ACC then is defined as

$$\mathrm{ACC}_l = \frac{1}{T}\sum_t^T \frac{\sum_i^I \sum_j^J w(i) f'_{t,l,i,j} o'_{t,i,j}}{\sqrt{\sum_i^I \sum_j^J w(i) {f'_{t,l,i,j}}^2 \sum_i^I \sum_j^J w(i) {o'_{t,i,j}}^2}} \tag{5}$$

The range of ACC values goes from 1, indicating perfect correlation, to -1, indicating perfect anti-correlation. A climatological forecast has an ACC value of zero. ECMWF states that when "ACC value falls below 0.6 it is considered that the positioning of synoptic scale features ceases to have value for forecasting purposes"[10].

4.3.3 Bias

The mean error, or simply bias, is computed for each location i, j as

$$\mathrm{Bias}_{l,i,j} = \frac{1}{T}\sum_t^T f_{t,l,i,j} - o_{t,i,j} \tag{6}$$

In addition, we compute the globally averaged root mean squared bias (RMSB) as

$$\mathrm{RMSB}_l = \sqrt{\frac{1}{IJ}\sum_i^I \sum_j^J w(i) \mathrm{Bias}^2_{l,i,j}} \tag{7}$$

4.3.4 Stable Equitable Error in Probability Space – SEEPS

Traditional deterministic scores such as RMSE and ACC are not good choices for evaluating precipitation forecasts. This is because precipitation has a very skewed distribution and high spatio-temporal intermittency or unpredictability. Under such conditions, traditional scores heavily favor unrealistically smooth forecasts. This is the case for all variables but is especially dramatic for skewed variables. For this reason, ECMWF and WMO decided to use the SEEPS score (Rodwell et al., 2010) for their routine deterministic precipitation evaluation. The SEEPS score is based on a three-class categorization into "dry", "light" and "heavy" precipitation. The score is designed to discourage "hedging" (i.e. smooth forecasts) and be stable to parameter choices. For details behind the score, please consult Rodwell et al. (2010). Here, we describe how the score is computed and how the computation differs slightly from that in the original paper.

We use a dry threshold of 0.25 mm/day. The remaining precipitation values are classified into light and heavy, so that climatologically there are twice as many light compared to heavy precipitation days. We compute this by calculating the 2/3rd quantile of non-dry days for each day-of-year according to the same procedure for computing a smooth climatology described in Section 3.2. This differs from most other SEEPS computations which uses monthly climatologies. Because the daily climatology is smooth though, this does not affect the results much. Forecast/observation pairs

[10] `https://confluence.ecmwf.int/display/FUG/Anomaly+Correlation+Coefficient`

are then categorized into the three classes and a 3x3 contingency table is created for each forecast lead time. The contingency table is then multiplied by the scoring matrix S based on the yearly average climatological occurrence of dry days p_1 for each geographical location:

$$S = \frac{1}{2} \begin{bmatrix} 0 & \frac{1}{1-p_1} & \frac{4}{1-p_1} \\ \frac{1}{p_1} & 0 & \frac{3}{1-p1} \\ \frac{1}{p_1} + \frac{3}{2+p_1} & \frac{3}{2+p_1} & 0 \end{bmatrix}$$

where columns represent observed probabilities and rows represent forecast probabilities.

Very wet and very dry regions are excluded. Here we use $0.01 < p_1 < 0.85$. The lower threshold is smaller than suggested in Rodwell et al. (2010). This is because we found that the suggested threshold of 0.1 would result many grid points being excluded. We think this is because we are using ERA5 as a ground truth which has more frequent light precipitation days ("drizzle") compared to observations. Finally an area-weighted average is taken over all locations.

4.4 Probabilistic metrics

4.4.1 Continuous ranked probability score (CRPS)

Given scalar ground truth Y, and i.i.d. predictions X, X', CRPS is defined as $\mathbb{E}|X-Y| - \frac{1}{2}\mathbb{E}|X-X'|$ (Gneiting and Raftery, 2007). The skill term, $\mathbb{E}|X-Y|$ penalizes poor predictions, while $-\frac{1}{2}\mathbb{E}|X-X'|$ encourages spread. CRPS is minimized just when X is drawn from the same distribution as Y. To see this, consider two i.i.d. ground truth samples, Y, Y', then subtract $\frac{1}{2}\mathbb{E}|Y-Y'|$ from CRPS to arrive at the divergence relation

$$\mathbb{E}\left|X-Y\right| - \frac{1}{2}\mathbb{E}\left|X-X'\right| - \frac{1}{2}\mathbb{E}\left|Y-Y'\right| \quad = \quad \int \left(\mathrm{P}[X \le y] - \mathrm{P}[Y \le y]\right)^2 \, dy.$$

Thus, scalar CRPS is equal, up to a constant independent of the prediction, to the squared L2 difference of their cumulative distribution functions. In the case of a deterministic prediction, the CRPS reduces to the MAE.

WeatherBench considers multi-dimensional predictions, $f_{t,l}$, conditional on the initial ground truth o_{t-l}. For these, CRPS is defined by averaging over time/latitude/longitude components. We also take advantage of $M \ge 2$ predictions $\{f^{(1)}, \dots, f^{(M)}\}$. Setting

$$\|g\|_{t,l} \quad := \quad \frac{1}{IJ} \sum_{i}^{I} \sum_{j}^{J} w(i) |g_{t,l,i,j}|,$$

we define

$$CRPS_l := \frac{1}{T} \sum_{t}^{T} \left[\frac{1}{M} \sum_{m=1}^{M} \|f^{(m)} - o\|_{t,l} - \frac{1}{2M(M-1)} \sum_{m=1}^{M} \sum_{n=1}^{M} \|f^{(m)} - f^{(n)}\|_{t,l} \right]. \tag{8}$$

This is the time average of unbiased (conditional) CRPS (Zamo and Naveau, 2018). With large enough T, this is an accurate estimate of CRPS conditional on o_{t-l}. It is therefore minimized by *any* prediction $f_{t,l}$ with the same component distributions as ground truth at time t, conditional on o_{t-l}.

The second term (8) is efficiently computed in $O(M \log M)$ time using $O(M)$ memory with a sort rather than a double summation. CRPS for deterministic predictions (where $M = 1$) is also supported, and in this case CRPS reduces to (weighted) mean absolute error.

4.4.2 Spread-skill ratio

The spread-skill ratio R is defined as the ratio between the ensemble spread and the RMSE of the ensemble mean $\overline{f}_{t,l,i,j} = \frac{1}{M}\sum_m^M f_{t,l,i,j,m}$.

$$\text{Spread}_l = \frac{1}{T}\sum_t^T \sqrt{\frac{1}{IJ}\sum_i^I\sum_j^J w(i) var_m(f_{t,l,i,j,m})} \tag{9}$$

with var_m being the variance in the ensemble dimension.

$$R = \frac{\text{Spread}}{\text{RMSE}(\overline{f})} \tag{10}$$

A well-calibrated ensemble forecast should have a spread-skill ratio of 1 (Fortin et al., 2014). Smaller values indicate an underdispersive forecast, while larger values indicate an overdispersive forecast. Note that the spread-skill ratio is only a first-order test for calibration. To further diagnose ensemble calibration, rank histograms are a suitable choice (Wilks, 2006, Ch. 7.7.2). We plan to include those in WB2 soon.

4.5 Energy Spectra

Zonal spectral energy along lines of constant latitude is computed as a function of wavenumber (unitless), frequency (m^{-1}) and wavelength (m).

With f_l discrete values along a zonal circle of constant latitude, with circumference C, the DFT F_k is computed as

$$F_k = \frac{1}{L}\sum_{l=0}^{L-1} f_l e^{-i2\pi kl/L}.$$

The energy spectrum is then set to

$$S_0 := C|F_0|^2, \qquad S_k = 2C|F_k|^2, \quad k = 1, \ldots, \lfloor L/2 \rfloor.$$

The factor of 2 appears for wavenumbers $k > 0$ since they account for both negative and positive frequency content.

This choice of normalization ensures that Parseval's relation for energy holds: Supposing f_l are sampled values of continuous function $f(\ell)$, for $0 < \ell < C$ (m), then (C/L) is the spacing of longitudinal samples, whence

$$\int_0^C |f(\ell)|^2\, d\ell \approx \frac{C}{L}\sum_{l=0}^{L-1} |f_l|^2 = \sum_{k=0}^{\lfloor L/2 \rfloor} S_k.$$

To arrive at a final spectrum we average the zonal spectra for 30° $< |lat| <$ 60°.

4.6 Headline scores

The headline scores are listed in Table 3. They reflect the most commonly evaluated variables in medium-range forecasting. The upper-level variables are chosen to capture the large-scale evolution of the atmosphere. Z500 and T850 are particularly good tracers of extra-tropical dynamics. Q700 provides a proxy for moisture transport and, indirectly, clouds. The surface variables are closely aligned with weather impact through 2 m temperature (T2M), 10 m wind speed (WS10) and 24 h precipitation accumulation (TP24hr). The mean sea level pressure (MSLP) is another measure of larger scale dynamics and is a good proxy for the strength of tropical and extra-tropical cyclones.

Variable	Short name	Deterministic metric	Probabilistic metric
		Upper-level variables	
500hPa Geopotential	Z500	RMSE	CRPS
850hPa Temperature	T850	RMSE	CRPS
700hPa Specific humidity	Q700	RMSE	CRPS
850hPa Wind vector/speed	WV/WS850	RMSE (WV)	CRPS (WS)
		Surface variables	
2m Temperature	T2M	RMSE	CRPS
10m Wind speed	WS10	RMSE	CRPS
Mean sea level pressure	MSLP	RMSE	CRPS
24h precipitation	TP24hr	SEEPS	CRPS

Table 3: List of headline scores.

5 Results

The description of the results in this paper will focus on the general characteristics of the evaluation metrics and the baselines, with a smaller focus on comparing the data-driven models. This is because the paper is a snapshot of the state-of-the-art of data-driven modeling at the time of writing and will almost certainly be outdated soon. For an up-to-date view of the data-driven state-of-the-art, visit the accompanying website (`https://sites.research.google/weatherbench`).

5.1 Headline scores

Fig. 1 shows a scorecard of the deterministic headline scores relative to the IFS HRES baseline. Figs. 2 and 3 show the absolute and relative deterministic scores as a function of lead time. Fig. 4 shows the CRPS for IFS ENS and the probabilistic climatology. Figures for ACC, bias and the spread-skill ratio can be found in the supplement.

On deterministic metrics (RMSE and ACC), IFS ENS (mean) preforms better than IFS HRES for longer lead times. For smaller scale variables this transition happens earlier. For example, for larger scale variables like pressure and temperature IFS ENS (mean) has lower errors up to around 2 days, compared to up to around 12h for wind and humidity. GraphCast and Pangu-Weather also score well compared to IFS HRES, with scores similar to those shown in the respective papers. Note that the Pangu-Weather scores shown here are slightly less skillful than those in the original paper. This is likely because the evaluation here is done on 2020, not 2018, but Pangu-Weather is only trained up to 2017. Lam et al. (2022) show that training on more recent years improves performance noticeably. Comparing IFS ENS (mean) to GraphCast, GraphCast has a lower error up to 3–6 days, after which the ensemble mean has the lowest error. This will be discussed in conjunction with the blurring of forecasts below.

It is also noticeable, especially in the plots relative to IFS HRES, that the scores for the data-driven methods have a strong 6h zig-zag pattern. This is an artifact of training and evaluating with ERA5, which has a 12h assimilation window. Forecasts initialized at 00/12UTC are initialized towards the beginning of the 09–21/21–09UTC ERA5 assimilation windows. This means that for the first 6h forecast, e.g., from 00 to 06 UTC, all the AI models have to learn is to emulate the IFS model version used in ERA5. For the next 6h window, e.g., from 06 to 12UTC, the models also have to learn the assimilation step—a harder task. If forecasts initialized at 06/18UTC were evaluated, the zig-zag pattern would be reversed. Note that in Lam et al. (2022), this pattern is not visible because evaluation is done on a 12h interval, thereby only sampling every second point

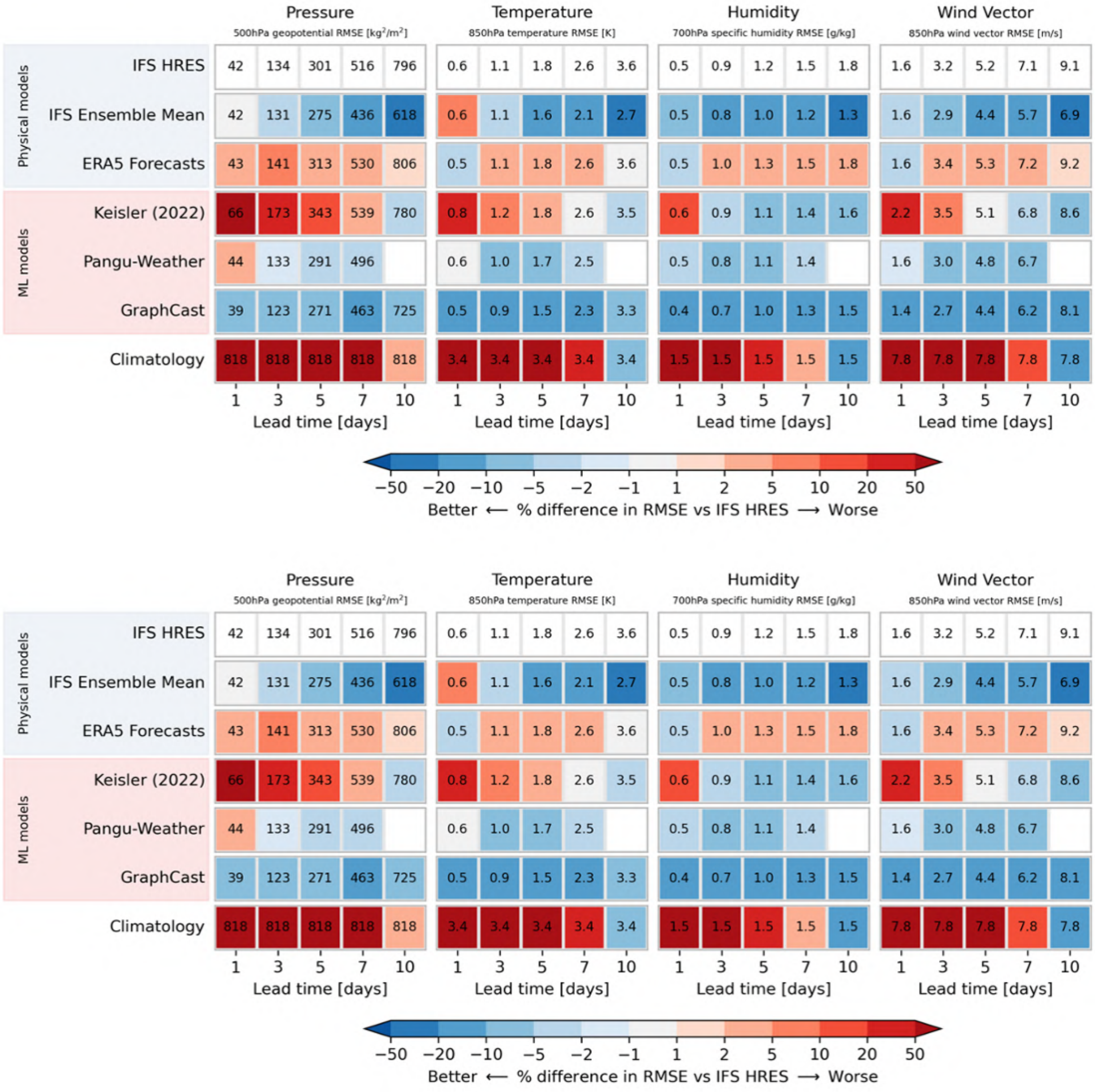

Figure 1: Deterministic headline scorecards for upper-level (top) and surface (bottom) variables. Values show absolute RMSE. Colors denote % difference to the IFS HRES baseline.

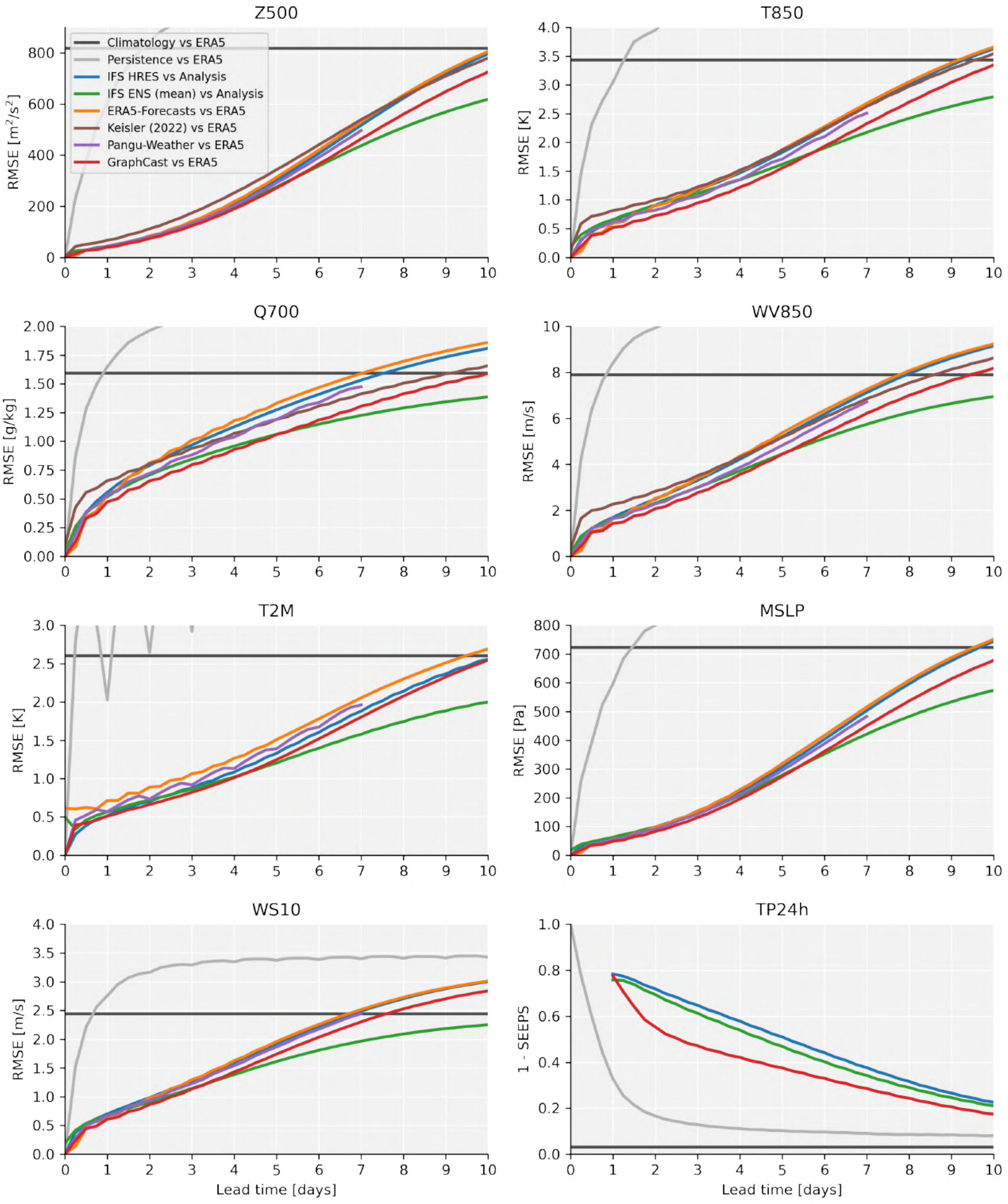

Figure 2: Global RMSE (SEEPS for TP24h) for headline variables for the year 2020. Note that for TP24h, IFS HRES and IFS ENS (mean) are evaluated against ERA5, since no precipitation accumulations are available for the analysis. Not all models/datasets have all variables available.

shown here. Here we chose to show the 6h evaluation for completeness.

Note that the larger error of IFS ENS (mean) at time $t = 0$ for 2m temperature stems from a difference in resolution between the ensemble (0.2°) and the analysis (0.1°). Results for the newly upgraded higher resolution ensemble (not included in WB2 yet) do not show this large initial error.

The precipitation verification using the SEEPS score shows that IFS HRES is the most skillful model with the "blurrier" models, such as IFS ENS (mean) and GraphCast, being less skillful. This shows the advantage of using a categorical score compared to an average score such as RMSE. On RMSE, the order of the models is reversed (not shown but also visible in ACC in Fig. S1).

At the time of writing, the probabilistic headline scores only contain two baselines, IFS ENS and the probabilistic climatology. The skill of IFS ENS approaches that of the probabilistic climatology towards the end of the two week forecast horizon. This is in line with the expected ∼15 day limit of predictability of synoptic weather (Selz, 2019). For some variables, while the IFS ENS skill curves flatten off, there is still a gap to the climatological skill. This could have several reasons: first, our methodology for computing the probabilistic climatology might not be a perfect representation of the true climatology; second; the forecast might still have some skill there. Especially for large-scale variables such as pressure and temperature, there is evidence of skill in the sub-seasonal range; and third, especially for temperature variables, climate change has a non-negligible effect over the 30 years used to compute the climatology.

An analysis of the spread-skill relationship (Supplement Figs. S4 and S5) shows that except for precipitation, which is underdispersive, the IFS ensemble forecast is very well calibrated, especially for longer lead times. The initial hours show some spin-up with more spread in the first hours, followed by a 2 day dip in the spread-skill ratio. For geopotential and pressure, this trend is reversed. Here it is important to not over-interpret global spread-skill ratios. More fine-grained analysis of tropical vs extra-tropical dispersion show more differentiated behavior. These results are included on the official website.

Fig. S3 also shows a time series of 6 day RMSE over Europe for all of 2020 for each of the models. Here, several forecast bust events are visible, where forecast skill drops significantly below normal (Rodwell et al., 2013). Interestingly, some bust cases are shared between the physical and AI models (e.g. around March 19) while for others AI models continue to perform well, despite IFS HRES and ENS performing badly (e.g. August 11). In some other cases, AI model spuriously perform badly. This is especially pronounced for Keisler (2022), while GraphCast and Pangu-Weather seem to be relatively robust, with only a few isolated "busts". Generally, this implies that AI models show similar error characteristics to physical models.

5.2 Bias

Fig. 5 shows the RMS Bias for the headline variables. Additionally, bias maps for 2m temperature, 24h precipitation, 10m wind speed and 700hPa specific humidity can be found in the supplement. It is important to note that one year is a small sample to compute bias statistics. In the bias maps in the supplement it is evident that dominant weather patterns in 2020 influence the bias. For example, 2020 saw a persistent heat wave in Siberia. All models tend to have a cold bias there since they fail to fully predict such a persistent pattern. Therefore, not too much should be interpreted into regional biases without considering the context.

What is noticeable is that while ML methods tend to have a lower average bias compared to physical methods, especially at early lead times, for wind speed Keisler (2022) and GraphCast show a large increase in average bias with lead time. The bias maps (Fig. S9) show that this is due to a consistent bias towards smaller wind speed values. This bias is not present when looking at U or V separately (see supplement of Lam et al. (2022)). What this indicates is that ML models trained

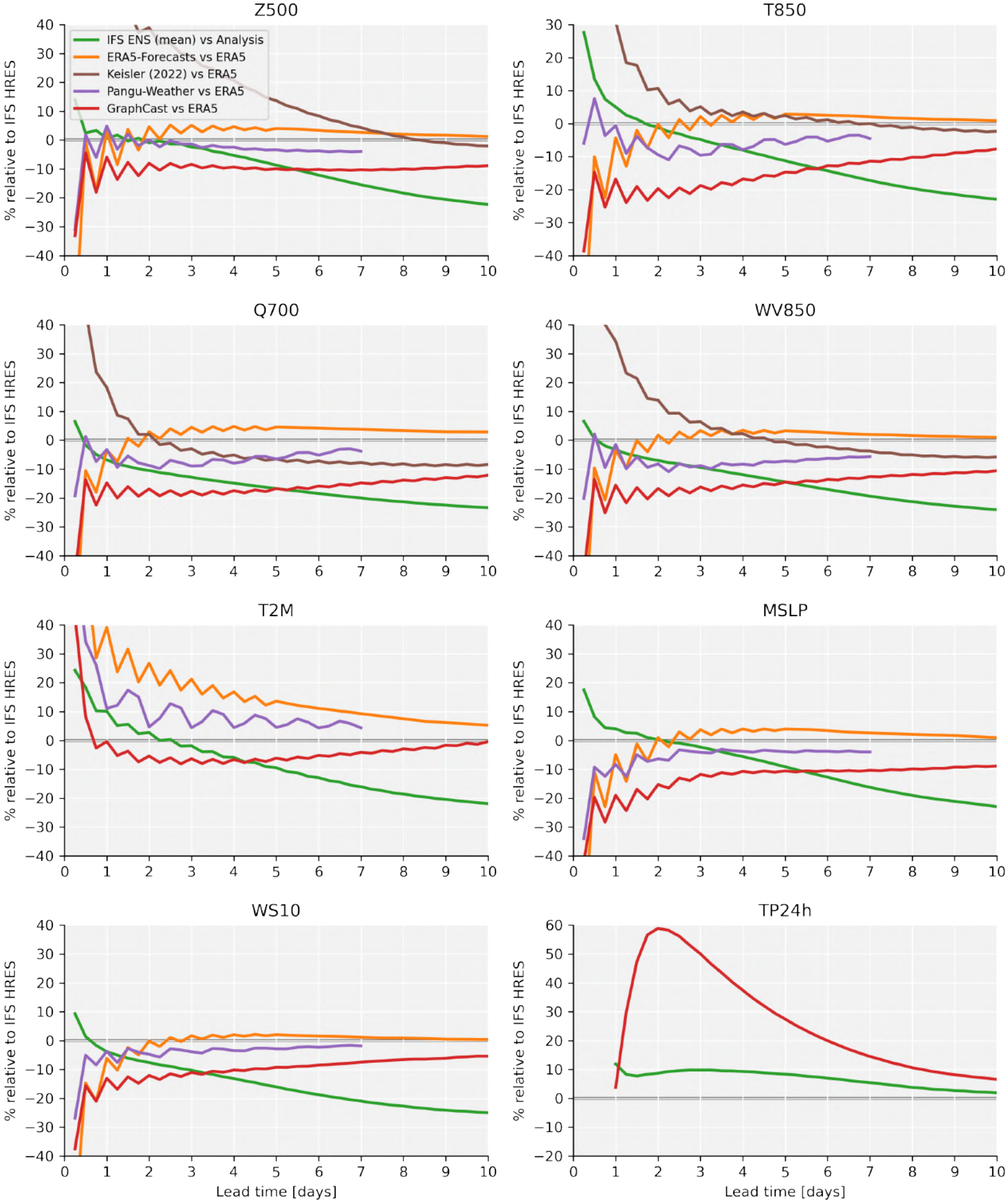

Figure 3: Global RMSE/SEEPS % difference compared to IFS HRES for headline variables for the year 2020. Negative values indicated lower RMSE.

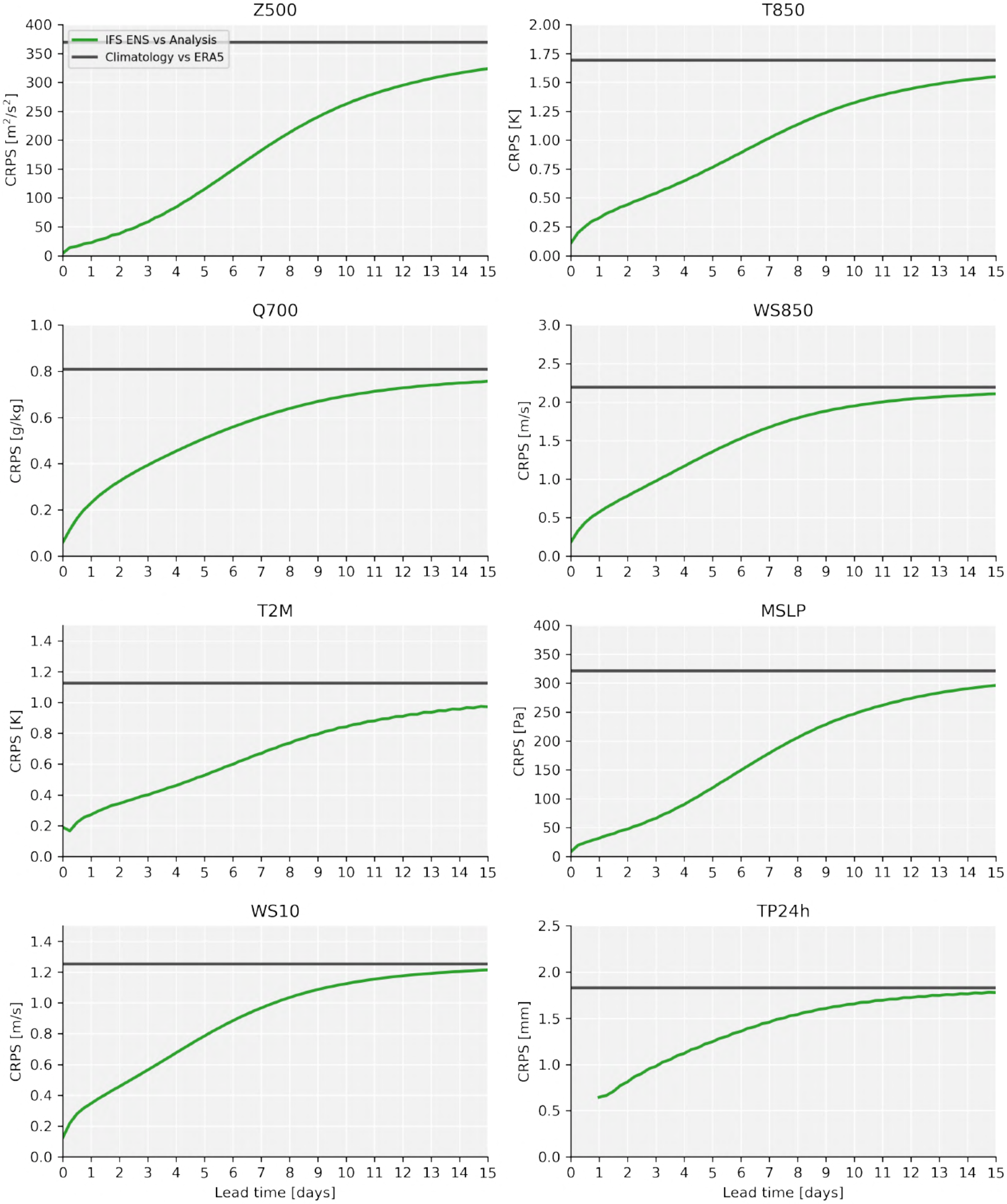

Figure 4: Global CRPS for headline variables for the year 2020.

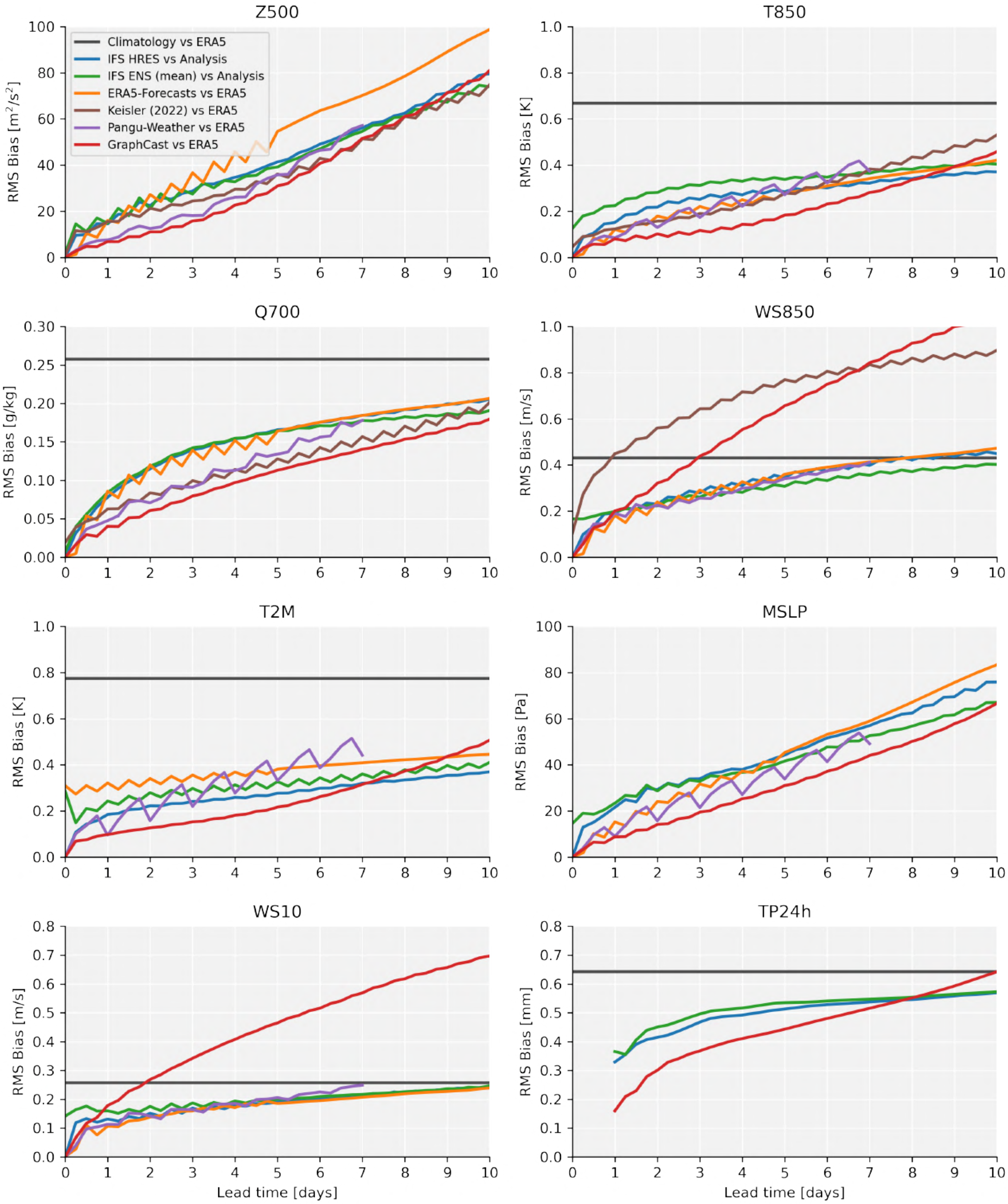

Figure 5: Global RMSB for the headline variables for the year 2020. For Z500 and MSLP the climatological bias is comparatively large and is not shown within the figure limits

to minimize each wind component separately, and doing a good job at doing so, can struggle to represent correlations between those components. Interestingly, Pangu-Weather does not have this bias.

5.3 Spectra

Fig. 6 shows the power spectra for four variables and four lead times. The ERA5 spectrum can be taken as a reference to compare between lead times as it is constant with lead time. IFS HRES generally has more energy on small scales compared to ERA5, likely due to its higher resolution but is also quasi-constant with lead time which is to be expected for a physical model. The IFS ENS mean shows an increasing drop-off in power with lead time, starting with the smaller wavelengths and then also for the larger wavelengths for longer lead times. Here it is important to re-iterate that the ensemble mean does not represent a realistic realization of weather but rather the mean of the forecast distribution. In other words, the mean will become increasingly smooth with lead time. The progression of the drop-off from smaller towards longer wavelengths is consistent with the model of upscale error growth first proposed by Lorenz (1969). For variables that vary more on small scales, such as humidity and precipitation, this effect is larger.

For Z500, the AI models tend to have more energy on small scales, more so for longer lead times. For Keisler (2022) this is clearly visible, e.g. in Fig. 7, where the geopotential contours have significant small scale "noise". For Pangu-Weather and GraphCast, the differences are hard to see visually. This increase in small scale energy for Z500 is in contrast to the behavior for the other variables where the AI models show a significant drop in small scale variability from 6h to 3d, similar to the ensemble mean. This is a result of the aforementioned smoothing. After 3 days the AI models tend to have a constant spectrum. This again is in line with the 24–72 h optimization windows. In other words, these models tend to optimize their blurring for those time ranges. The ensemble mean continues to become blurrier resulting in lower RMSE values for longer lead times. The Pangu-Weather spectra also show several wiggles and spikes. These could be caused by the tiling of the transformer patches.

The spectra clearly show some of the blurring exhibited by AI models. It is important to note though that having a spectrum that matches observations well is a necessary but not sufficient conditions for "realism".

5.4 Case studies

While case studies can only provide anecdotal evidence of model performance, they offer a more "human" and holistic view that skill scores cannot provide. Here we shows one such case study, with three more in the supplement for the readers' own inspection. On the WB2 website, we plan to extend this catalog of case studies.

Fig. 7 shows model forecasts and the corresponding ERA5 ground truth for storm Alex which brought damaging wind and rain across Europe.[11]. What is noticeable is that all models, physical and AI-based, predict the evolution of the storm with impressive accuracy and agreement up to four days lead time. October 2 saw strong winds over the North-West of France as well as parts of northern Spain. IFS HRES and even the ENS mean reasonably predict this, even though not as far inland as in ERA5. The AI models also show large values of wind speed. Keisler (2022) and GraphCast fail to develop this feature and instead show a trough stretching from Iceland to Britain.

[11] https://climate.copernicus.eu/esotc/2020/storm-alex

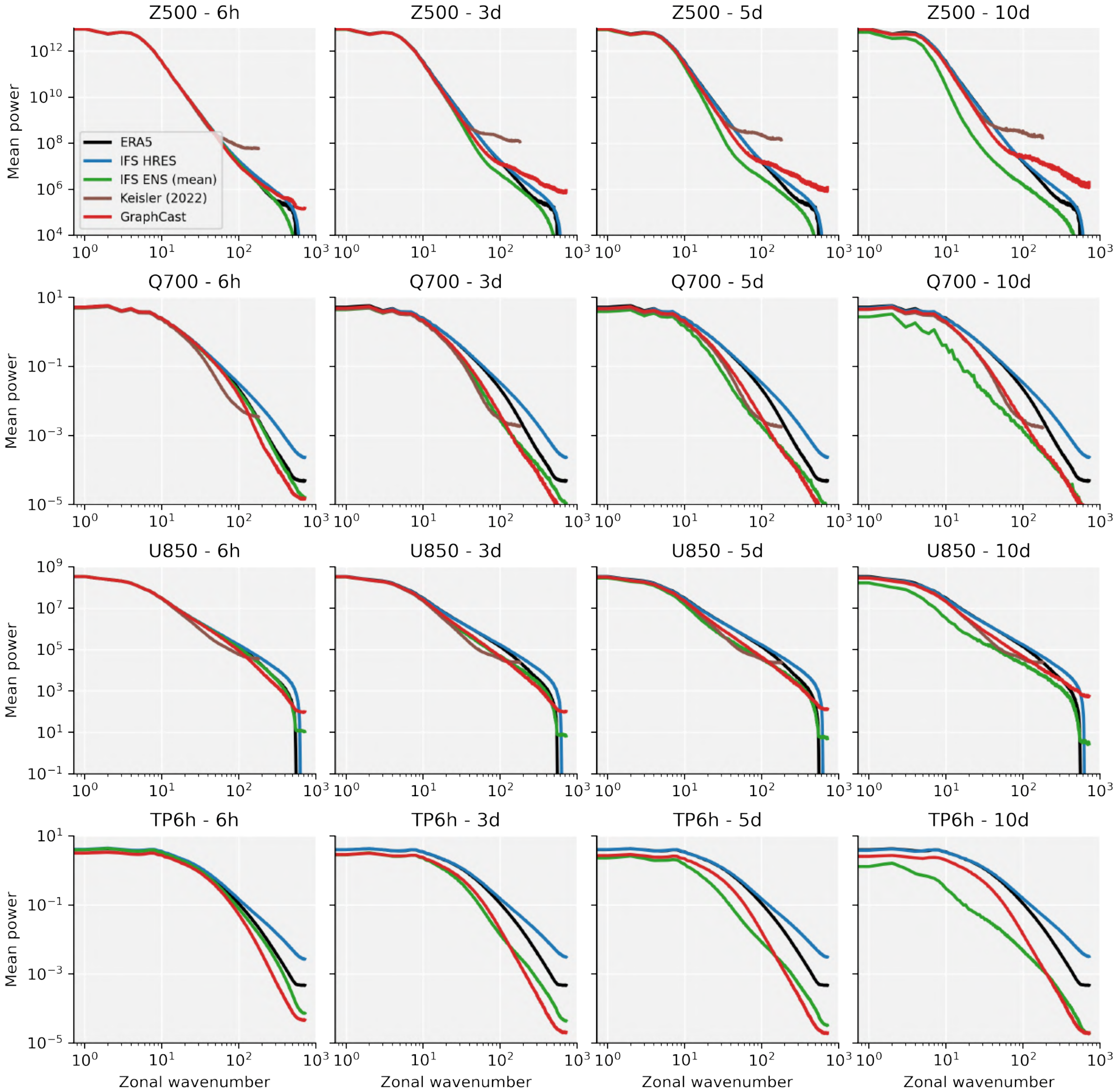

Figure 6: Power spectra of 500hPa geopotential, 700hPa specific humidity, 850hPa u-wind and 6h precipitation accumulation for different lead times.

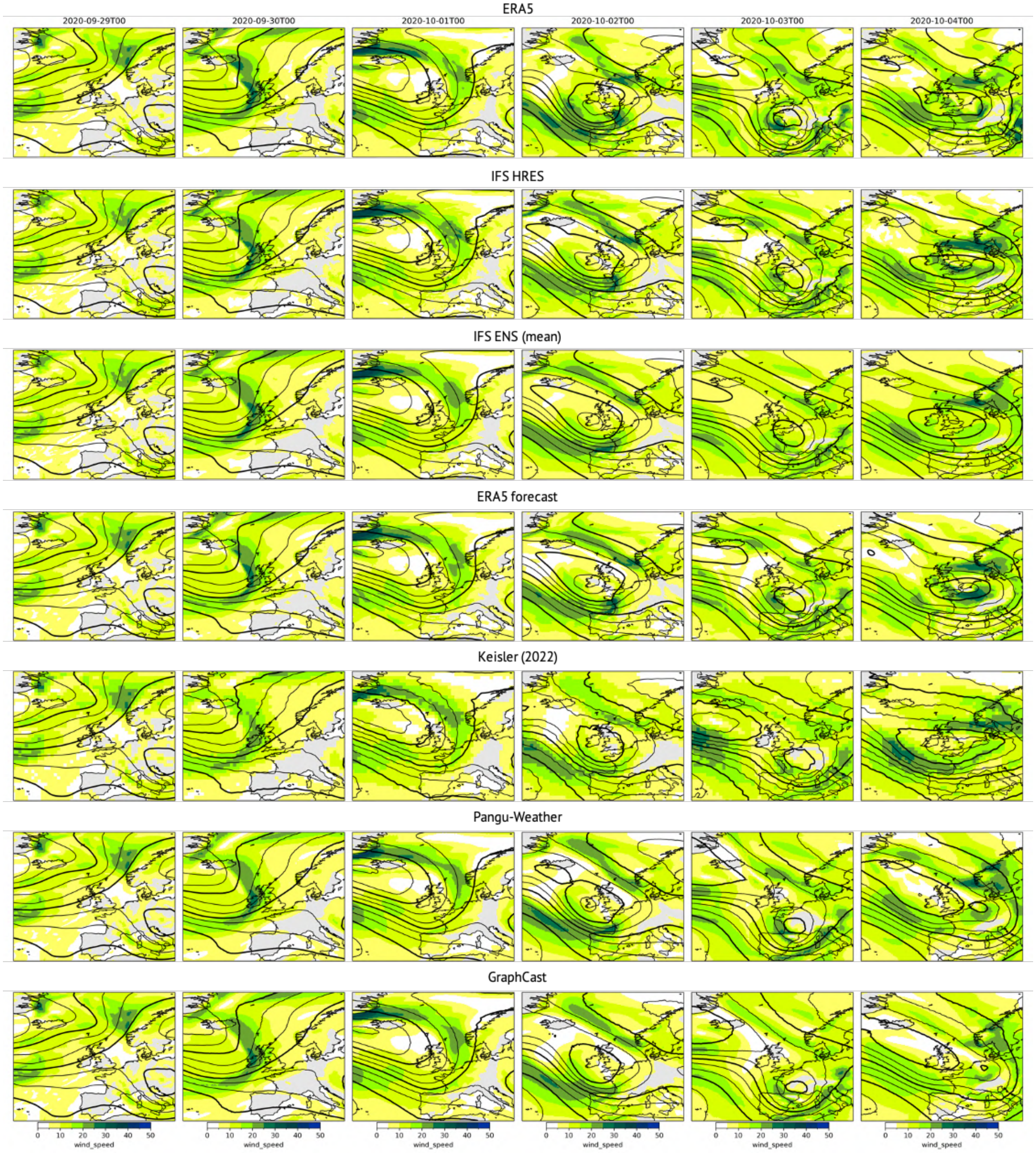

Figure 7: Case study: Storm Alex. Top row shows ERA5 ground truth. Other rows show forecasts initialized at 2020-09-29 00UTC. Lines show Z500 contours. Shading shows 850hPa wind speed.

Supp. Fig. S14 shows the evolution of Hurricane Laura, the most damaging hurricane of the 2020 season.[12] IFS HRES shows impressive skill in predicting the evolution and track of the hurricane up to 5 days in advance with the timing and location of landfall on August 8 almost perfectly forecasted. The IFS ENS mean somewhat predictably fails to predict the strength and location of the hurricane because it averages out several hurricane tracks. Interestingly the ERA5 forecasts, while decent in forecasting the track, struggle to predict the intensity of the hurricane, perhaps a consequence of the lower resolution of the model. Keisler (2022) has a track that is too far West. Pangu-Weather and GraphCast have a solid track forecast with reasonable cyclone structure but fail to predict the intensity in pressure and wind speed seen in the actual hurricane of the IFS forecast.

We plan to add more case studies to the website in the future. While case studies should not be over-interpreted, they are evidence that while sometimes lacking in detail and fidelity, AI models can produce skillful extreme weather forecasts for events outside of their training dataset.

6 Discussion

6.1 ERA5 as ground truth

Here, we use ERA5 as our ground truth dataset. However, as already discussed in Section 3.1, ERA5 is a model simulation that is kept close to observations, rather than direct observations. The quality of ERA5 depends on the variable in question. Large-scale dynamical variables like geopotential and upper-levels temperatures and wind tend to be well represented in reanalysis datasets. For surface variables, the sparsity of observations and difficulty of representing small-scale physics in the underlying model can cause larger discrepancies with observations. This is especially true for precipitation, which is not directly assimilated into ERA5 (e.g., through radar observations) and often show large differences to rain gauges or radar precipitation estimates (see e.g. Lavers et al., 2022; Andrychowicz et al., 2023). The precipitation evaluation using ERA5 shown here should really be seen as a placeholder for more accurate precipitation data.

Operational weather services like ECMWF verify their forecasts with direct observations, e.g., from weather stations, in addition to using assimilated ground truths. This is something we are looking to add to WeatherBench in the future. However, station observations come with their own issues: first, they are unevenly distributed with some regions being especially sparse in observations; second, station data comes in varying quality and requires careful quality control; and third, station data suffers from representativeness issues; that is, the station might not be representative of the model grid box it is compared to.

For these reasons, ERA5 still represents a good option for evaluation for most variables because of its temporal and spatial coverage but it should be kept in mind that it does not tell the full story. Notably, some regional ML and post-processing models already look at using data directly for training and evaluation (Andrychowicz et al., 2023; Demaeyer et al., 2023).

6.2 ERA5 initialization versus operational forecasting

Most data-driven methods so far have been trained and evaluated using ERA5 data as initial conditions. As already discussed in Section 3.1, ERA5 data would not be available in real time to initialize live, operational forecasts. For a true apples-to-apples comparison with operational physical models, ML forecasts would need to be initialized using operationally available initial conditions. The first such experiment has recently been done by Ben-Bouallegue et al. (2023a), who took the

[12] `https://www.nhc.noaa.gov/data/tcr/AL132020_Laura.pdf`

Pangu-Weather model trained on ERA5 and initialized it with operational IFS analyses. For their 2022/23 evaluation period, the "operational" Pangu-Weather forecasts performed equally well or better than the ERA5 initialized forecasts, suggesting that the jump from ERA5 to operational initialization is manageable. Regardless, going forward it will be important to distinguish between ERA5-initialized and (quasi-)operational models as e.g. done in Table 1.

6.3 Forecast realism, probabilistic forecasts and extremes

One feature of current AI methods is that they produce unrealistically smooth forecasts, a trend that often becomes more pronounced with lead time. This is a direct consequence of these models minimizing a deterministic mean error (such as MSE). Because the evolution of the atmosphere is chaotic and forecast uncertainty grows with time, this will lead models to "hedge their bets" by predicting the mean of the potential forecast distribution. In a non-linear, high-dimensional system such as the atmosphere, the mean of many states is not a realistic state itself. The result are blurry forecasts that perform well on some skill metrics (like RMSE or ACC) but do not represent a possible weather state. This is evident in several analyses presented here: spectra show excessive blurring of AI models for longer lead times; the SEEPS score shows how blurry models (IFS ENS mean and GraphCast) fail to predict the right precipitation category; wind speed biases are evidence that current AI models have difficulties learning correlations between variables; and the case studies show that the intensity of local features such as wind speed, precipitation and cyclone intensity aren't represented. For some applications, having good forecasts of average weather is sufficient but for others AI models are not yet appropriate.

This is naturally related to probabilistic prediction, i.e., predicting the full range of potential weather outcomes, which is so important for predicting the probability of extreme weather. Current, "traditional" forecast systems use ensembles for this purpose. AI methods could follow a similar approach by producing an ensemble of generative roll-outs, or they could directly predict probabilistic outcomes of interest, such as the probability distribution of precipitation as done in the MetNet family of models (Andrychowicz et al., 2023) or many post-processing models (e.g. Gneiting et al. (2005)). For many applications, such as predicting weather at a particular location, the latter approach might be more straightforward and sufficient, while for other applications, for example cyclone track forecasting or when humans interpret model output, temporal and spatial structure is important. Working closely with end users will be key in determining the most appropriate probabilistic representation for each application.

The probabilistic evaluation metrics proposed in WB2 (CRPS and spread/skill) are very much just the tip of the iceberg. They are univariate statistics, i.e., they ignore spatial and temporal correlations, and they do not specifically focus on extreme weather. Just like with the deterministic metrics (RMSE or ACC), better CRPS values do not automatically mean a more useful forecast. Weather forecasting is a very high-dimensional problem, which means that there won't be a single metric to determine forecast quality. Rather, evaluation will have to be guided by applications.

6.4 Post-processing

In this paper, the focus has been on data-driven forecast models. However, AI can also and has for a long time been used to post-process dynamical forecasts (Rasp and Lerch, 2018; Finn, 2021; Gneiting and Raftery, 2007; Grönquist et al., 2021). WB2 can equally well serve as a benchmark for post-processing models, on top of dynamical forecasts such as those produced by ECMWF. In fact, comparing "purely" data-driven forecasting models with dynamical models with state-of-the-art post-processing should be an insightful exercise going forward. Note that benchmarks for post-processing have been proposed by Ashkboos et al. (2022) including re-forecasts for training post-processing models and an extreme weighted CRPS score definition. Similarly, Demaeyer et al. (2023) present a benchmark for station-based post-processing.

Significant improvements over raw dynamical model output can be expected using post-processing. Previous studies (e.g. Rasp and Lerch, 2018; Ben-Bouallegue et al., 2023b; Finn, 2021) suggest that probabilistic forecasts can be improved by up to 20% in terms of CRPS, depending on the variable in question. Since data-driven methods, at least partly, already perform a post-processing implicitly, a fair comparison would be against post-processed dynamical models. Another interesting question is how much AI models can benefit from additional post-processing. Hopefully, these comparisons will be added to WB2 soon.

7 Conclusion

WeatherBench 2.0 is an updated benchmark for data-driven, global weather forecasting. It is motivated by the rapid advances in used AI methods since the publication of the original WeatherBench benchmark. WB2 is designed to stay close to the operational forecast evaluation used by many weather centers and to provide a robust framework for evaluating new methods against operational baselines. By providing evaluation code and data, we hope to speed up machine learning workflows and ensure the reproducibility of results.

WB2 is also designed to be a dynamic framework that will be updated with new metrics and models as this area of research evolves. Several possible extensions have already been discussed in this paper, for example including station observations and evaluating extremes.

WB2 can be expanded with additional metrics and baselines depending on the needs of the community. Several possible extensions have already been discussed in the paper, particularly an increased focus on evaluating extremes and impact variables at fine scales, possibly by using station observations.

Acknowledgments

We would like to thank Kaifeng Bi from the Pangu-Weather team and Ryan Keisler for their help in adding their models to WeatherBench 2. Further, we thanks John Anderson and Nal Kalchbrenner for helpful discussions and comments, as well as Kunal Shah and Rahul Mahrsee for their help downloading the data.

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:2306.06079 [physics].

Ashkboos, S., Huang, L., Dryden, N., Ben-Nun, T., Dueben, P., Gianinazzi, L., Kummer, L., and Hoefler, T. (2022). ENS-10: A Dataset For Post-Processing Ensemble Weather Forecasts. arXiv:2206.14786 [physics].

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55.

Ben-Bouallegue, Z., Clare, M. C. A., Magnusson, L., Gascon, E., Maier-Gerber, M., Janousek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T. K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., and Pappenberger, F. (2023a). The rise of data-driven weather forecasting. arXiv:2307.10128 [physics].

Ben-Bouallegue, Z., Weyn, J. A., Clare, M. C. A., Dramsch, J., Dueben, P., and Chantry, M. (2023b). Improving medium-range ensemble weather forecasts with hierarchical ensemble transformers. arXiv:2303.17195 [physics].

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q. (2023). Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, pages 1–6.

Buizza, R., Milleer, M., and Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125(560):2887–2908. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49712556006.

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., Chen, X., Ma, L., Zhang, T., Su, R., Ci, Y., Li, B., Yang, X., and Ouyang, W. (2023a). FengWu: Pushing the Skillful Global Medium-range Weather Forecast beyond 10 Days Lead. arXiv:2304.02948 [physics].

Chen, L., Du, F., Hu, Y., Wang, F., and Wang, Z. (2023b). SwinRDM: Integrate SwinRNN with Diffusion Model towards High-Resolution and High-Quality Weather Forecasting. arXiv:2306.03110 [physics].

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., and Li, H. (2023c). FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. arXiv:2306.12873 [physics].

Clare, M. C., Jamil, O., and Morcrette, C. J. (2021). Combining distribution-based neural networks to predict weather forecast probabilities. *Quarterly Journal of the Royal Meteorological Society*, 147(741):4337–4357.

Demaeyer, J., Bhend, J., Lerch, S., Primo, C., Van Schaeybroeck, B., Atencia, A., Ben Bouallègue, Z., Chen, J., Dabernig, M., Evans, G., Faganeli Pucer, J., Hooper, B., Horat, N., Jobst, D., Merše, J., Mlakar, P., Möller, A., Mestre, O., Taillardat, M., and Vannitsem, S. (2023). The EUPPBench postprocessing benchmark dataset v1.0. *Earth System Science Data Discussions*, pages 1–25. Publisher: Copernicus GmbH.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. ISSN: 1063-6919.

Dueben, P. D. and Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11(10):3999–4009. Publisher: Copernicus GmbH.

Dueben, P. D., Schultz, M. G., Chantry, M., Gagne, D. J., Hall, D. M., and McGovern, A. (2022). Challenges and Benchmark Datasets for Machine Learning in the Atmospheric Sciences: Definition, Status, and Outlook. *Artificial Intelligence for the Earth Systems*, 1(3):e210002.

Finn, T. S. (2021). Self-Attentive Ensemble Transformer: Representing Ensemble Interactions in Neural Networks for Earth System Models. arXiv:2106.13924 [physics].

Fortin, V., Abaza, M., Anctil, F., and Turcotte, R. (2014). Why Should Ensemble Spread Match the RMSE of the Ensemble Mean? *Journal of Hydrometeorology*, 15(4):1708–1713.

Garg, S., Rasp, S., and Thuerey, N. (2022). WeatherBench Probability: A benchmark dataset for probabilistic medium-range weather forecasting along with deep learning baseline models. arXiv:2205.00865 [physics].

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214506000001437.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. (2005). Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, 133(5):1098–1118.

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., and Hoefler, T. (2021). Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194):20200092. Publisher: Royal Society.

Guibas, J., Mardani, M., Li, Z., Tao, A., Anandkumar, A., and Catanzaro, B. (2022). Adaptive Fourier Neural Operators: Efficient Token Mixers for Transformers. arXiv:2111.13587 [cs].

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Jung, T. and Leutbecher, M. (2008). Scale-dependent verification of ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 134(633):973–984. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.255.

Kalnay, E. (2002). *Atmospheric Modeling, Data Assimilation and Predictability.* Cambridge University Press, 1 edition.

Keisler, R. (2022). Forecasting Global Weather with Graph Neural Networks. arXiv:2202.07575 [physics].

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., Ravuri, S., Ewalds, T., Alet, F., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Stott, J., Vinyals, O., Mohamed, S., and Battaglia, P. (2022). GraphCast: Learning skillful medium-range global weather forecasting. arXiv:2212.12794 [physics].

Lavers, D. A., Simmons, A., Vamborg, F., and Rodwell, M. J. (2022). An evaluation of ERA5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society*, 148(748):3152–3165. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.4351.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20(2):130–141. Publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences.

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus*, 21(3):289–307.

Magnusson, L. and Källén, E. (2013). Factors Influencing Skill Improvements in the ECMWF Forecasting System. *Monthly Weather Review*, 141(9):3142–3153. Publisher: American Meteorological Society Section: Monthly Weather Review.

Palmer, T., Molteni, F., Mureau, R., Buizza, R., Chapelet, P., and Tribbia, J. (1993). Ensemble prediction. In *Palmer, T. N., et al. "Ensemble prediction." Proc. ECMWF Seminar on Validation of models over Europe. Vol. 1.*

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., and Anandkumar, A. (2022). FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. arXiv:2202.11214 [physics].

Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A., and Battaglia, P. W. (2021). Learning Mesh-Based Simulation with Graph Networks. arXiv:2010.03409 [cs].

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N. (2020). WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11).

Rasp, S. and Lerch, S. (2018). Neural Networks for Postprocessing Ensemble Weather Forecasts. *Monthly Weather Review*, 146(11):3885–3900.

Rasp, S. and Thuerey, N. (2021). Data-Driven Medium-Range Weather Prediction With a Resnet Pretrained on Climate Simulations: A New Model for WeatherBench. *Journal of Advances in Modeling Earth Systems*, 13(2):e2020MS002405. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002405.

Rodwell, M. J., Magnusson, L., Bauer, P., Bechtold, P., Bonavita, M., Cardinali, C., Diamantakis, M., Earnshaw, P., Garcia-Mendez, A., Isaksen, L., Källén, E., Klocke, D., Lopez, P., McNally, T., Persson, A., Prates, F., and Wedi, N. (2013). Characteristics of Occasional Poor Medium-Range Weather Forecasts for Europe. *Bulletin of the American Meteorological Society*, 94(9):1393–1405. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

Rodwell, M. J., Richardson, D. S., Hewson, T. D., and Haiden, T. (2010). A new equitable score suitable for verifying precipitation in numerical weather prediction: New Equitable Score for Precipitation in NWP. *Quarterly Journal of the Royal Meteorological Society*, 136(650):1344–1363.

Scher, S. (2018). Toward Data-Driven Weather and Climate Forecasting: Approximating a Simple General Circulation Model With Deep Learning. *Geophysical Research Letters*, 45(22):12,616–12,622. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018GL080704.

Selz, T. (2019). Estimating the Intrinsic Limit of Predictability Using a Stochastic Convection Scheme. *Journal of the Atmospheric Sciences*, 76(3):757–765.

Selz, T., Riemer, M., and Craig, G. C. (2022). The Transition from Practical to Intrinsic Predictability of Midlatitude Weather. *Journal of the Atmospheric Sciences*, 79(8):2013–2030.

Stensrud, D. J. (2007). *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models*. Cambridge University Press, 1 edition.

Toth, Z. and Kalnay, E. (1993). Ensemble Forecasting at NMC: The Generation of Perturbations. *Bulletin of the American Meteorological Society*, 74(12):2317–2330.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Weyn, J. A., Durran, D. R., and Caruana, R. (2019). Can Machines Learn to Predict Weather? Using Deep Learning to Predict Gridded 500-hPa Geopotential Height From Historical Weather Data. *Journal of Advances in Modeling Earth Systems*, 11(8):2680–2693.

Weyn, J. A., Durran, D. R., and Caruana, R. (2020). Improving Data-Driven Global Weather Prediction Using Deep Convolutional Neural Networks on a Cubed Sphere. *Journal of Advances in Modeling Earth Systems*, 12(9):e2020MS002109. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002109.

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences*. Number v. 91 in International geophysics series. Academic Press, Amsterdam ; Boston, 2nd ed edition.

World Meteorological Organization (2019). *Manual on the Global Data-processing and Forecasting System (WMO-No. 485): Annex IV to the WMO Technical Regulations*. WMO. WMO, Geneva, updated in 2021 edition.

Zamo, M. and Naveau, P. (2018). Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts. *Mathematical Geosciences*, 50(2):209–234.

Zhang, F., Bei, N., Rotunno, R., Snyder, C., and Epifanio, C. C. (2007). Mesoscale Predictability of Moist Baroclinic Waves: Convection-Permitting Experiments and Multistage Error Growth Dynamics. *Journal of the Atmospheric Sciences*, 64(10):3579–3594.

Zhang, F., Sun, Y. Q., Magnusson, L., Buizza, R., Lin, S.-J., Chen, J.-H., and Emanuel, K. (2019). What Is the Predictability Limit of Midlatitude Weather? *Journal of the Atmospheric Sciences*, 76(4):1077–1091.

Supplement

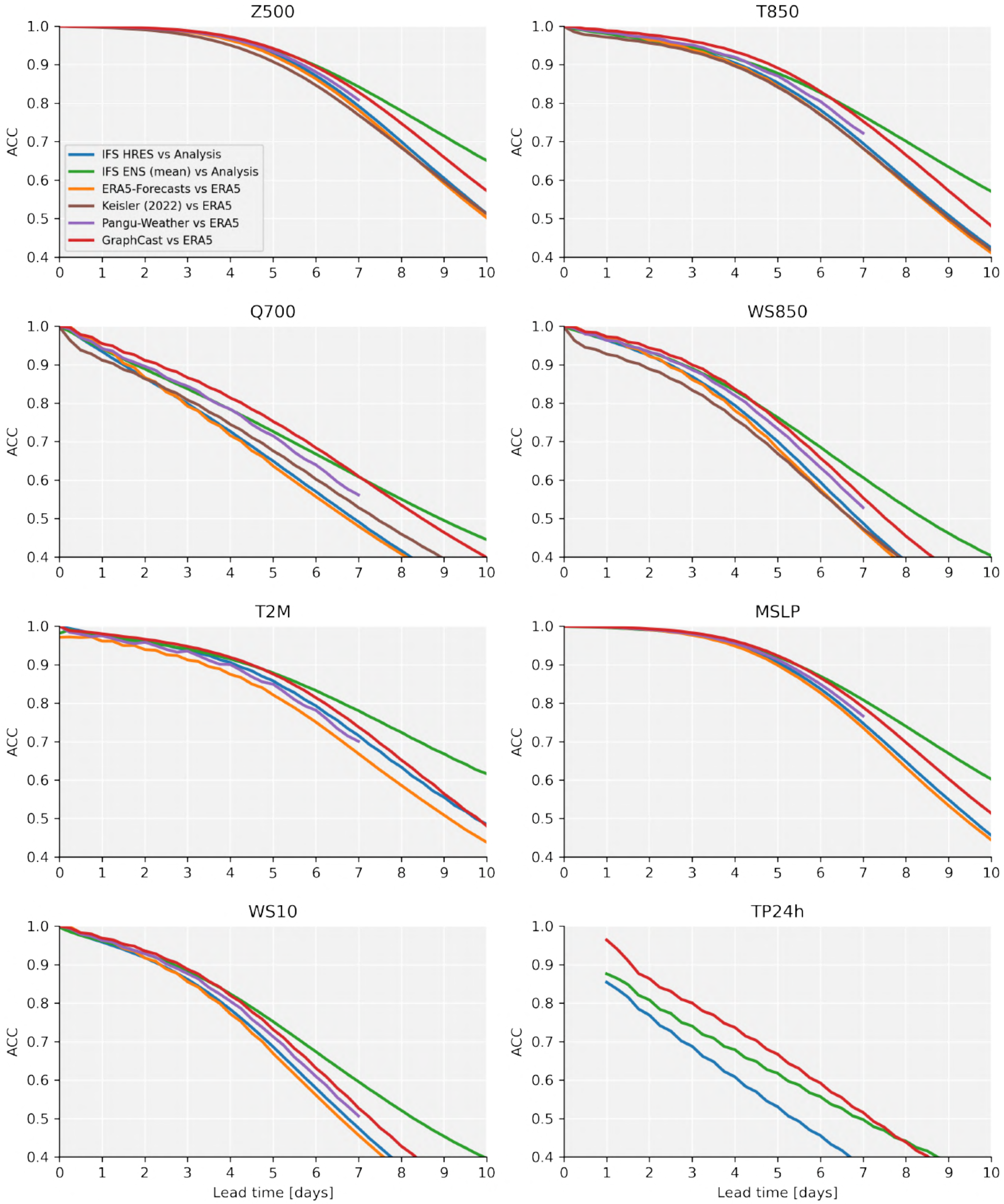

Figure S1: Global ACC for headline variables for the year 2020.

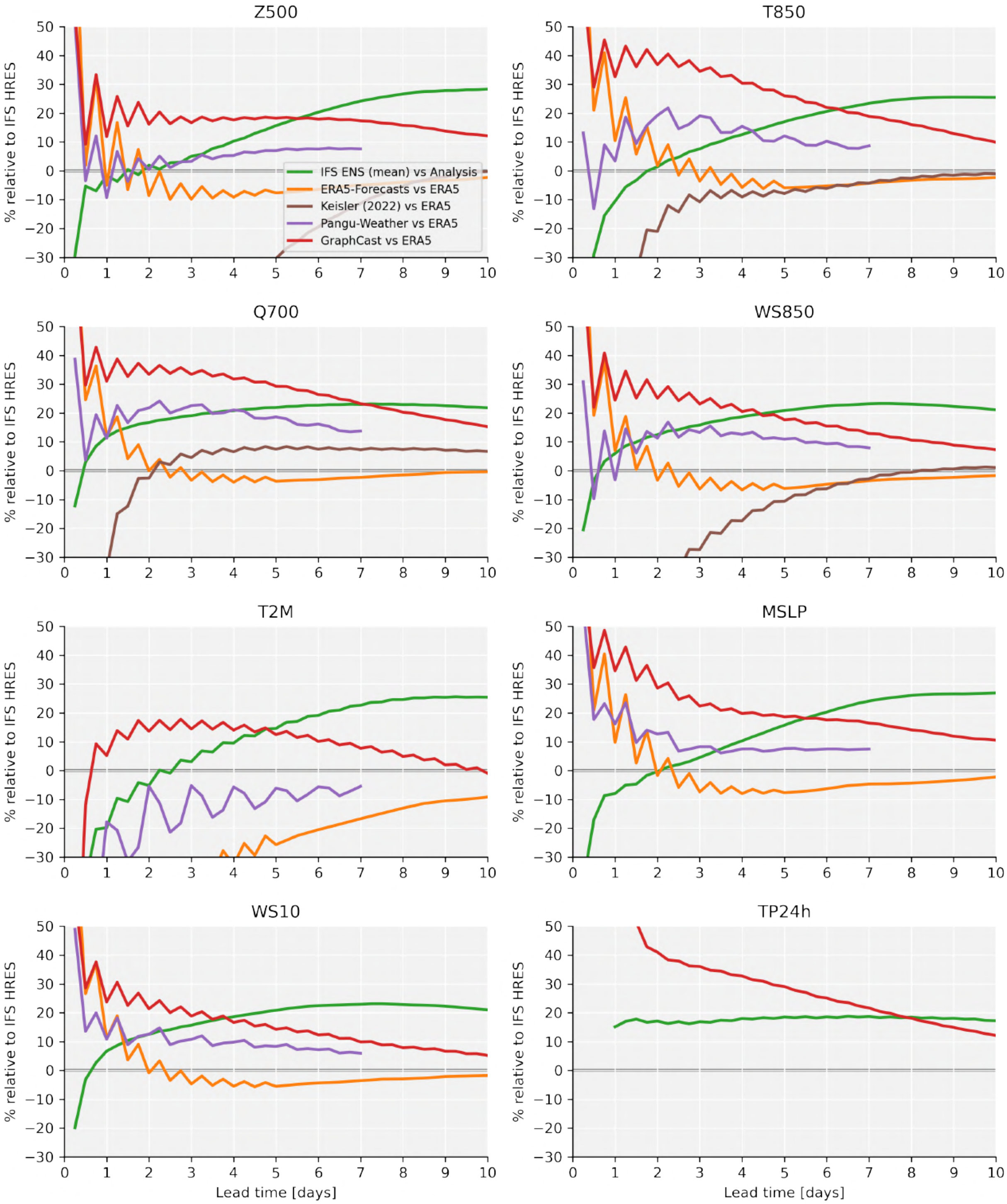

Figure S2: ACC on headline variables. % difference compared to IFS HRES. Positive values indicate higher ACC.

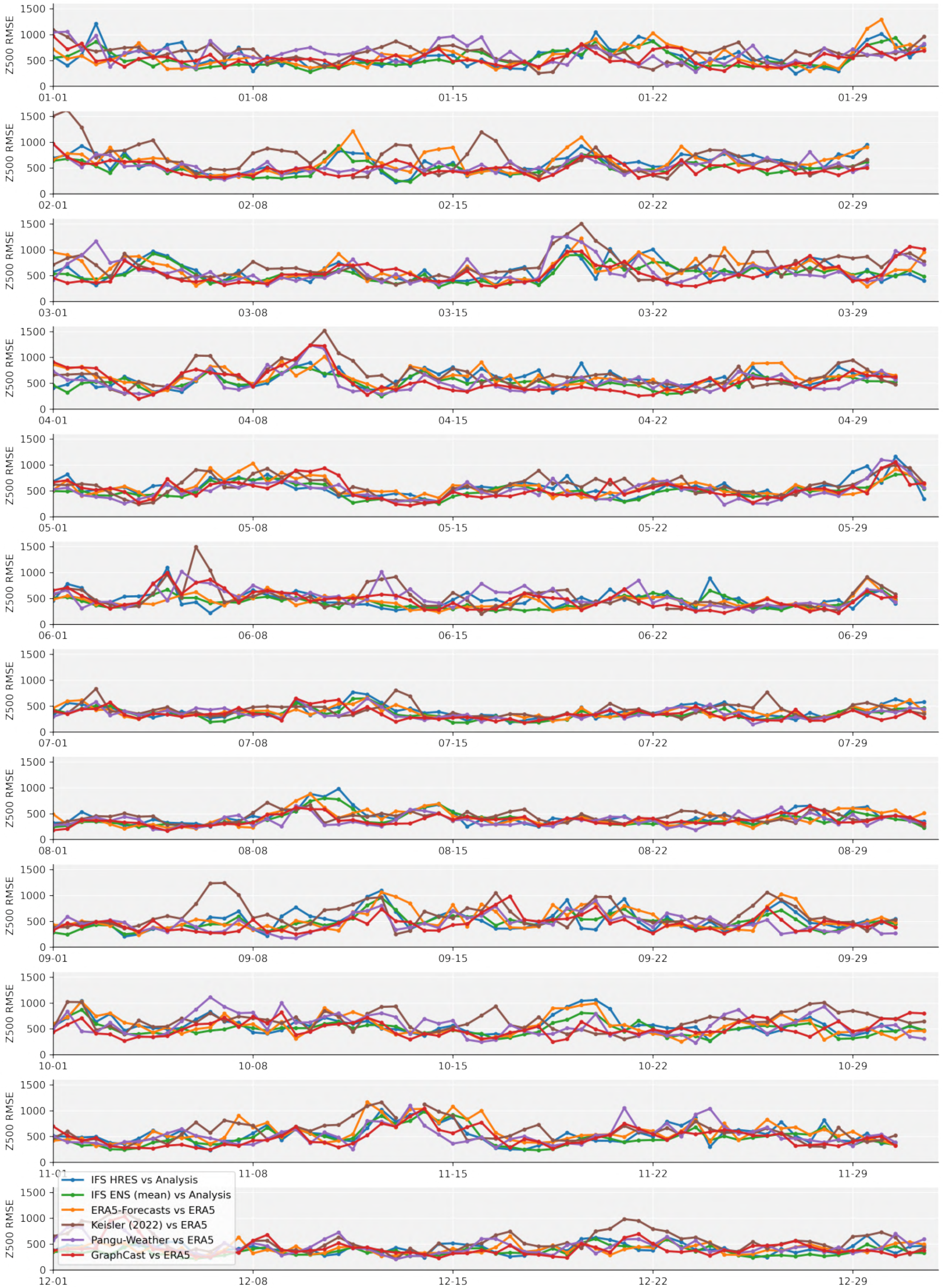

Figure S3: Timeseries of day 6 RMSE in m^2/s^2 over Europe defined by $35° < lat < 75°$ and $-12.5° < lon < 42.5°$. All models evaluated at 1.5° resolution. Climatological errors are omitted.

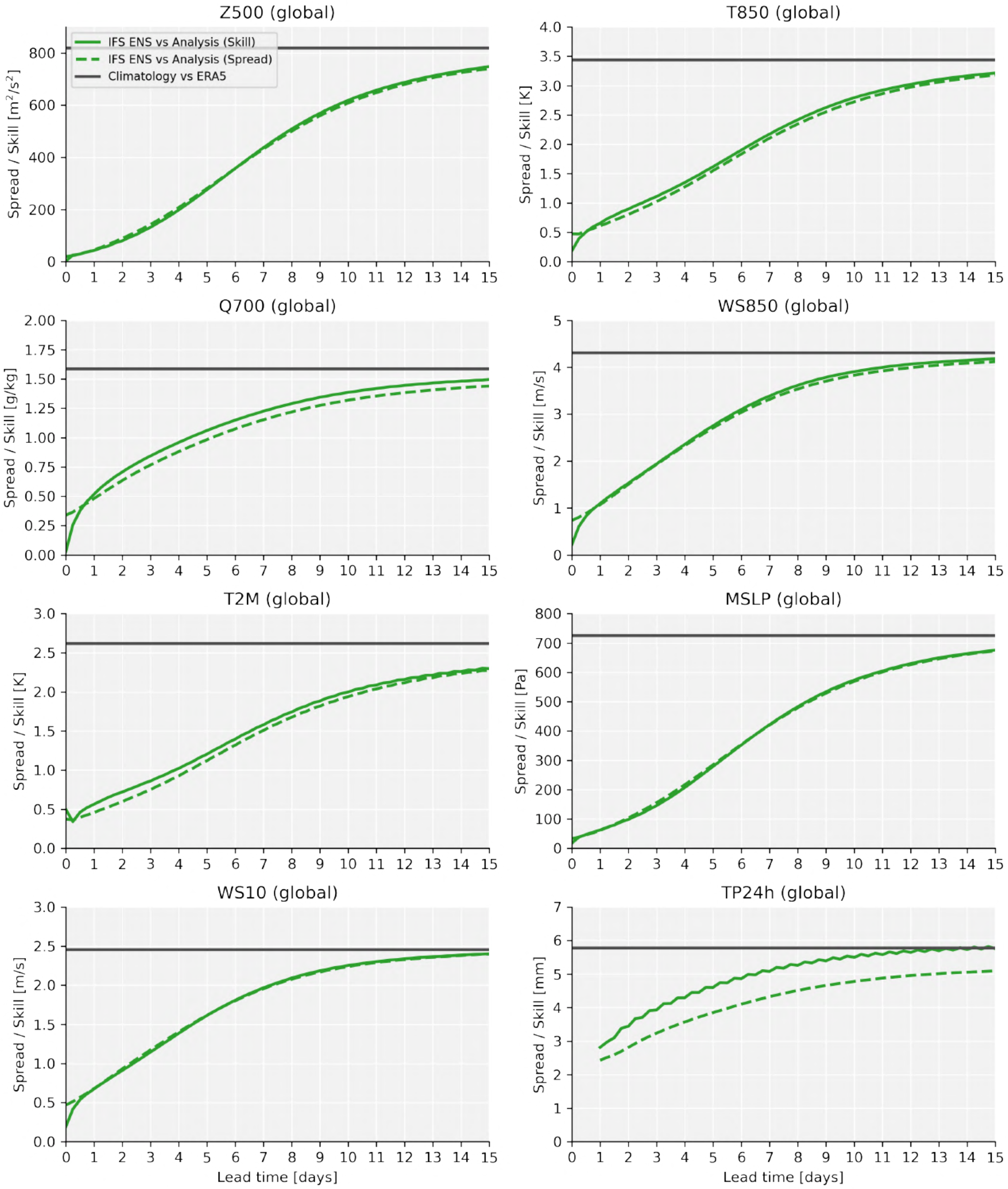

Figure S4: Spread (dashed) and skill (solid) for headline variables for the year 2020.

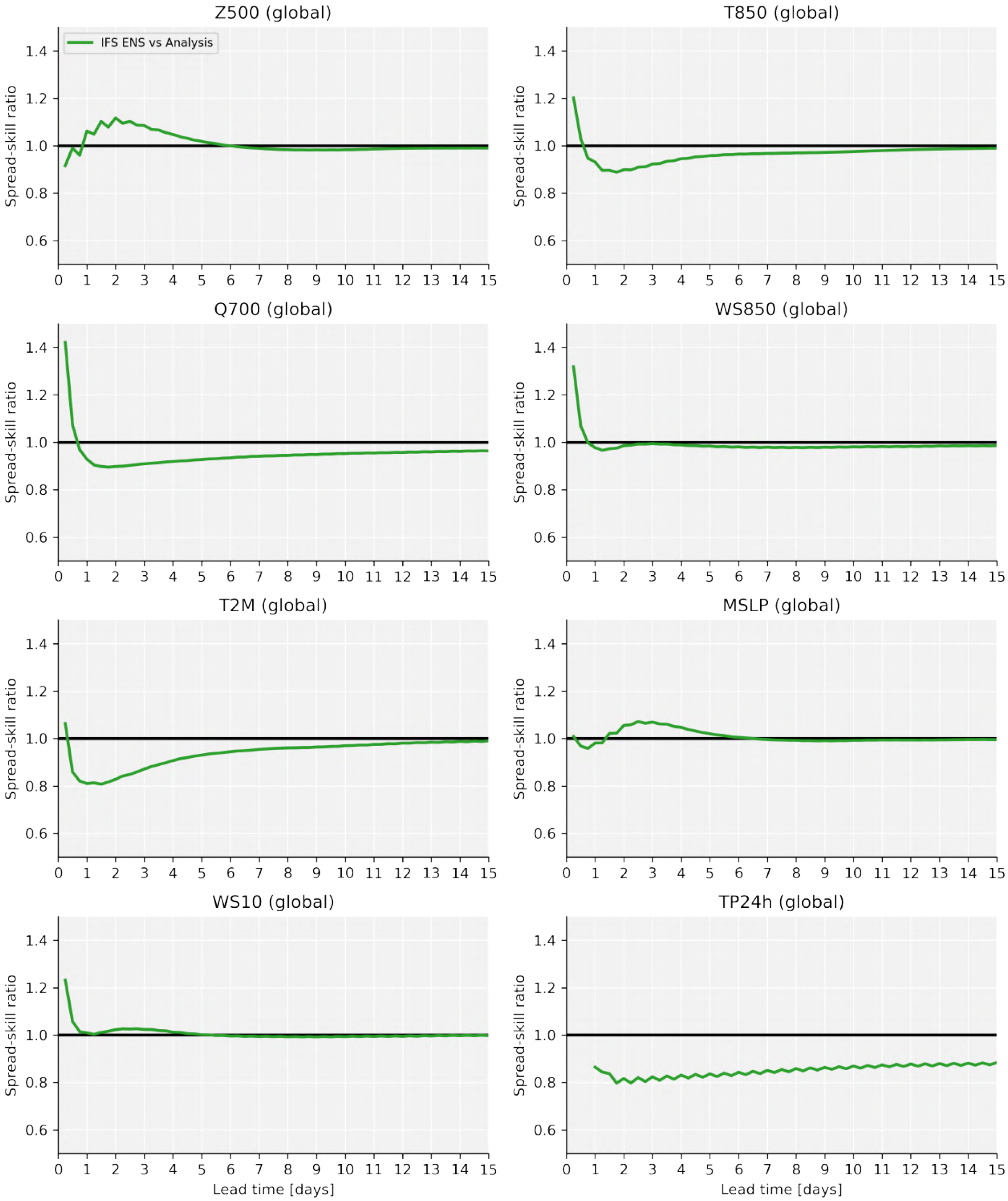

Figure S5: Spread/skill ratio for headline variables for the year 2020.

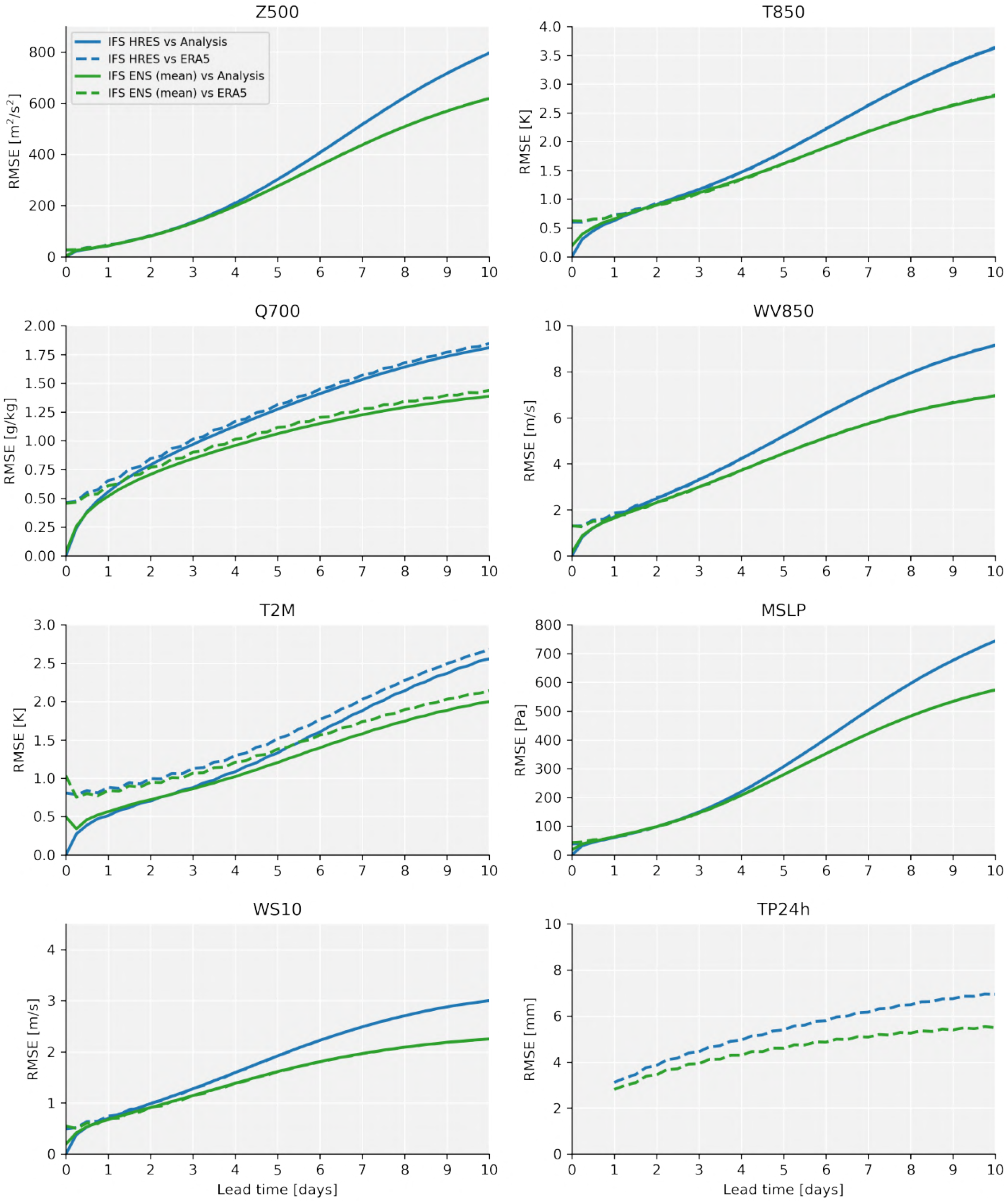

Figure S6: Comparison of evaluating IFS HRES and ENS (mean) with analysis and ERA5 ground truth.

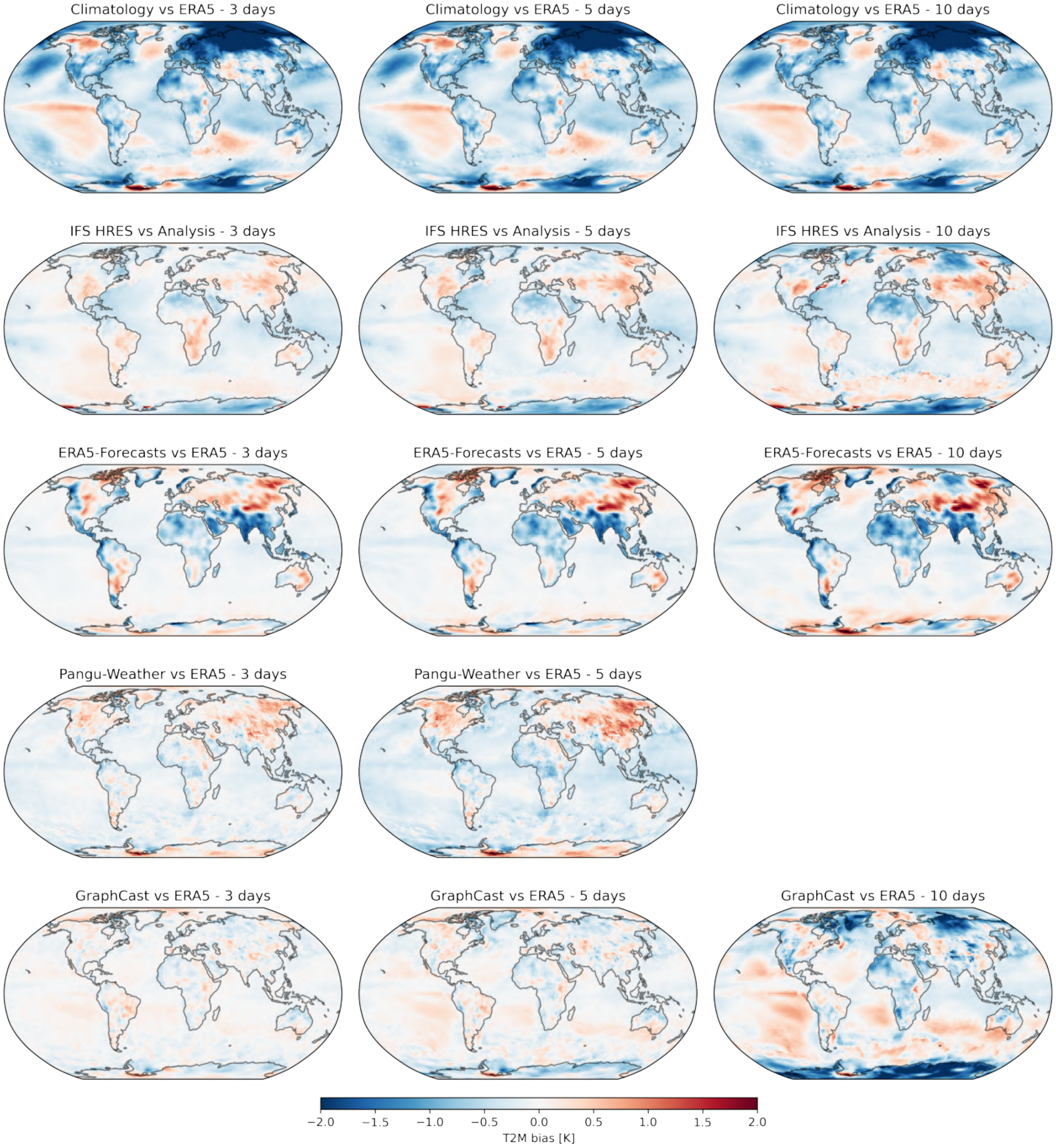
Climatology vs ERA5 - 3 days
Climatology vs ERA5 - 5 days
Climatology vs ERA5 - 10 days
IFS HRES vs Analysis - 3 days
IFS HRES vs Analysis - 5 days
IFS HRES vs Analysis - 10 days
ERA5-Forecasts vs ERA5 - 3 days
ERA5-Forecasts vs ERA5 - 5 days
ERA5-Forecasts vs ERA5 - 10 days
Pangu-Weather vs ERA5 - 3 days
Pangu-Weather vs ERA5 - 5 days
GraphCast vs ERA5 - 3 days
GraphCast vs ERA5 - 5 days
GraphCast vs ERA5 - 10 days
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
T2M bias [K]

Figure S7: Global mean 2m temperature bias for 3, 5 and 10 day lead times.

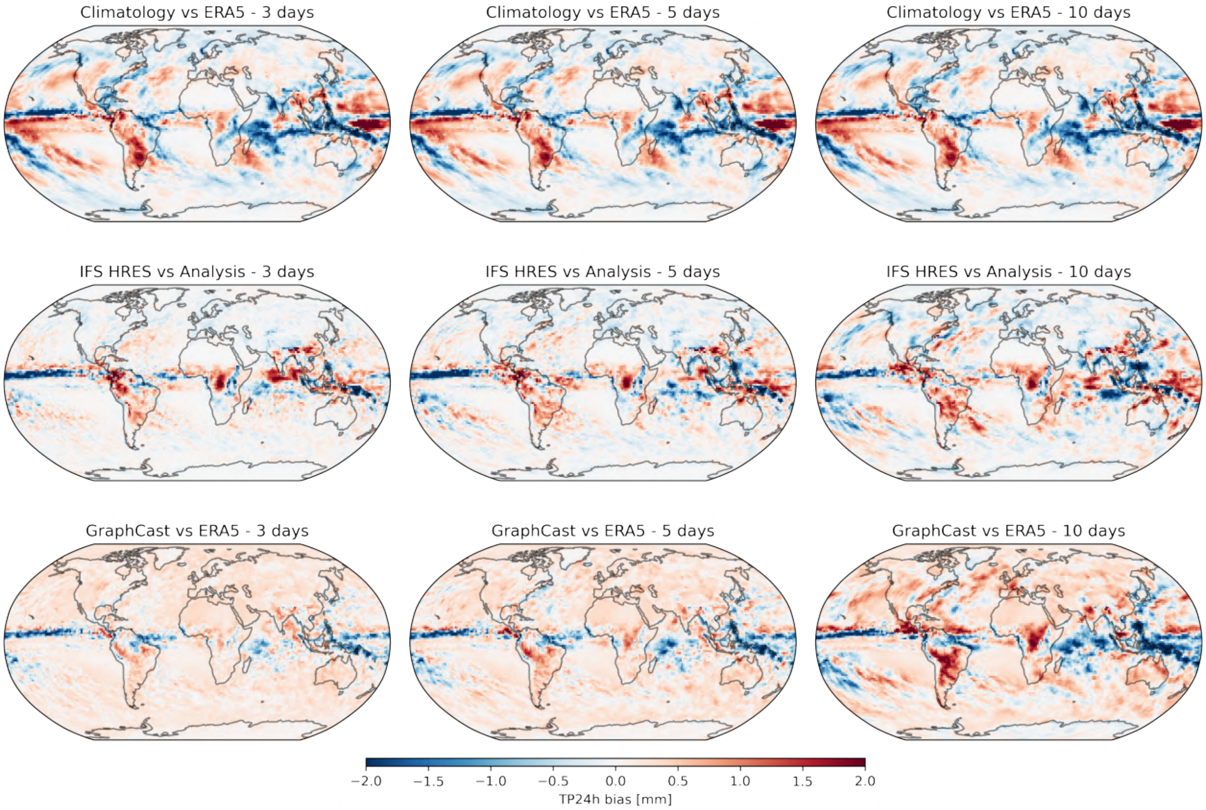
Climatology vs ERA5 - 3 days
Climatology vs ERA5 - 5 days
Climatology vs ERA5 - 10 days
IFS HRES vs Analysis - 3 days
IFS HRES vs Analysis - 5 days
IFS HRES vs Analysis - 10 days
GraphCast vs ERA5 - 3 days
GraphCast vs ERA5 - 5 days
GraphCast vs ERA5 - 10 days
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
TP24h bias [mm]

Figure S8: Global mean 24h precipitation accumulation bias for 3, 5 and 10 day lead times.

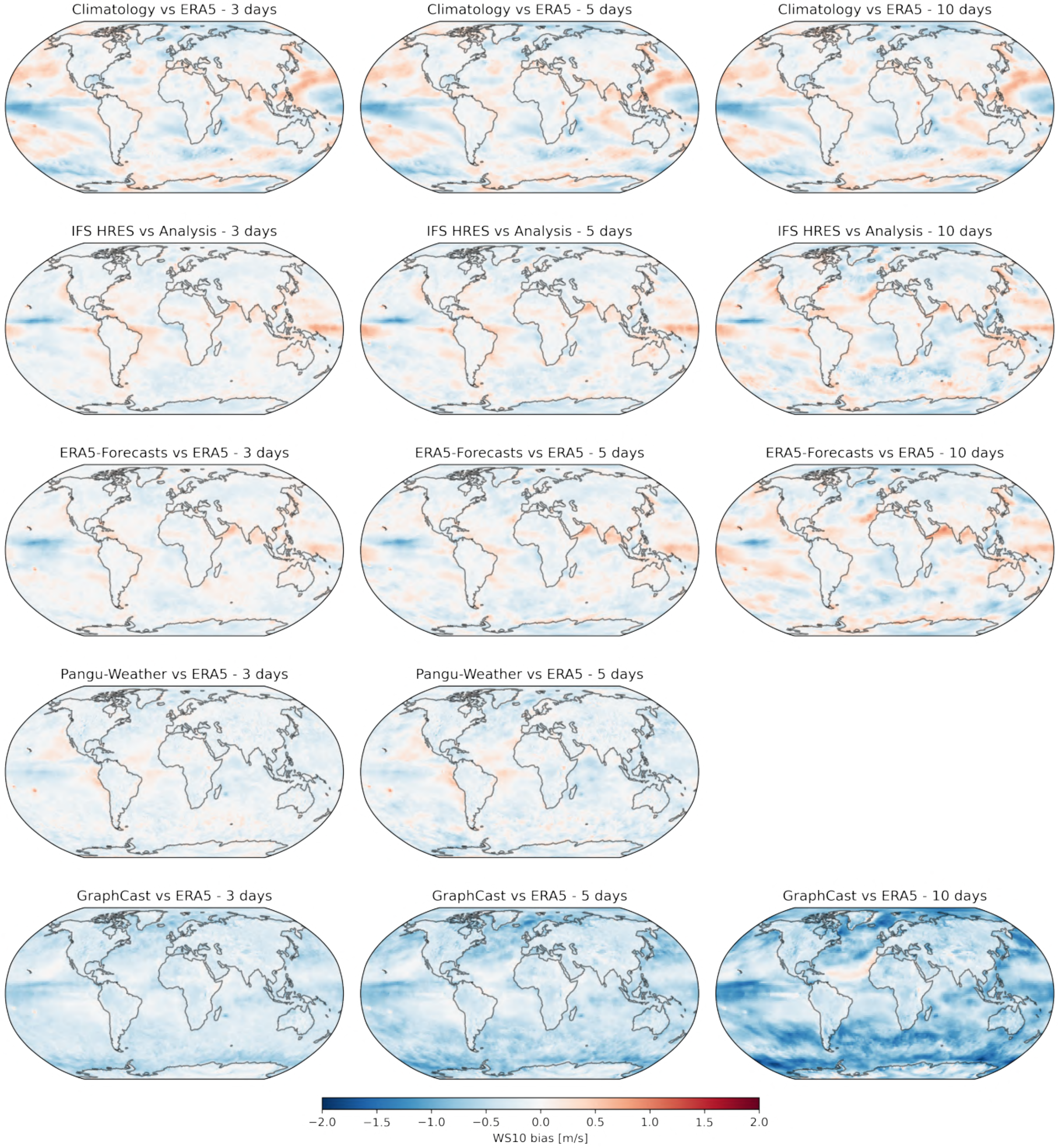
Climatology vs ERA5 - 3 days
Climatology vs ERA5 - 5 days
Climatology vs ERA5 - 10 days
IFS HRES vs Analysis - 3 days
IFS HRES vs Analysis - 5 days
IFS HRES vs Analysis - 10 days
ERA5-Forecasts vs ERA5 - 3 days
ERA5-Forecasts vs ERA5 - 5 days
ERA5-Forecasts vs ERA5 - 10 days
Pangu-Weather vs ERA5 - 3 days
Pangu-Weather vs ERA5 - 5 days
GraphCast vs ERA5 - 3 days
GraphCast vs ERA5 - 5 days
GraphCast vs ERA5 - 10 days
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
WS10 bias [m/s]

Figure S9: Global mean 10m wind speed bias for 3, 5 and 10 day lead times.

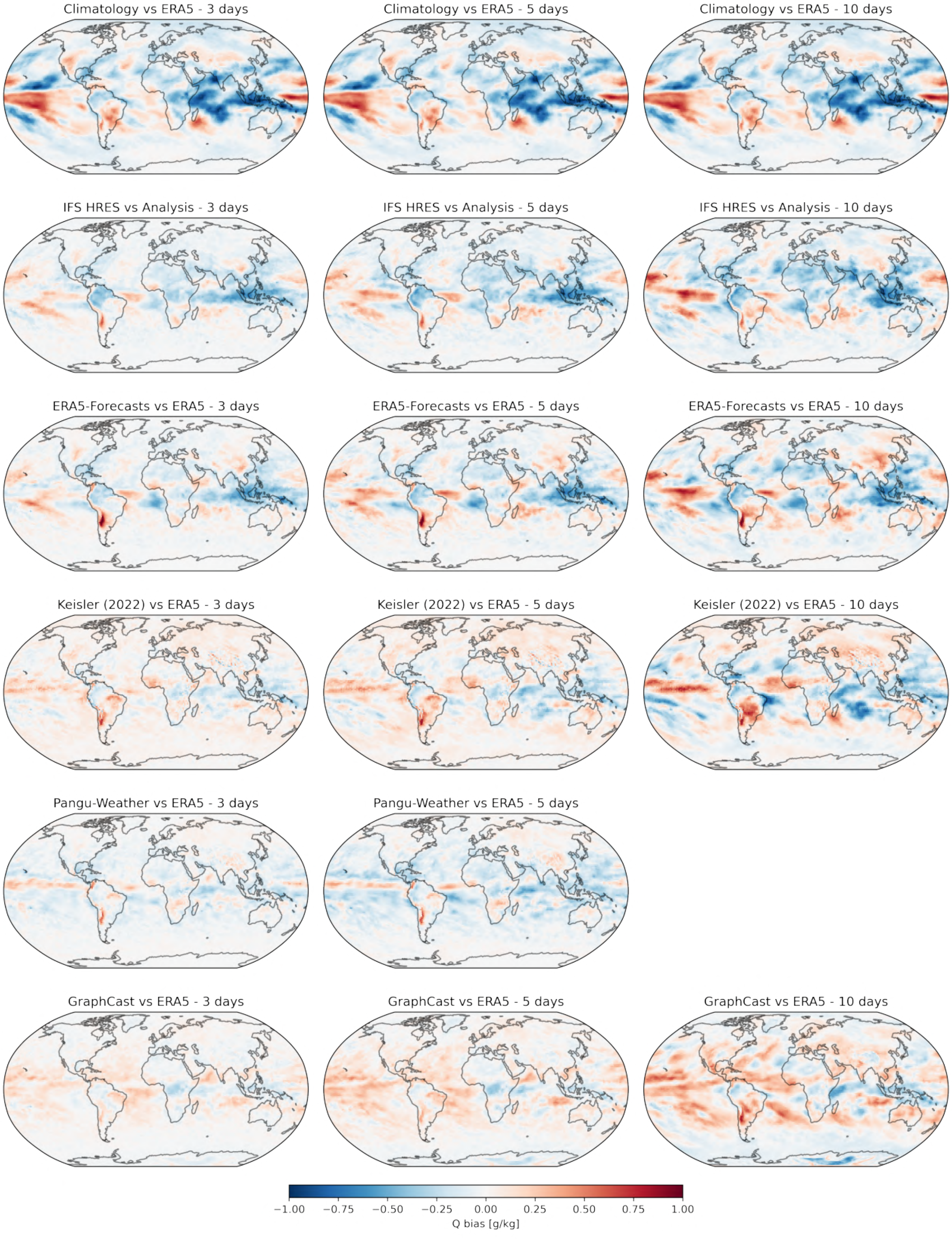
Climatology vs ERA5 - 3 days
Climatology vs ERA5 - 5 days
Climatology vs ERA5 - 10 days
IFS HRES vs Analysis - 3 days
IFS HRES vs Analysis - 5 days
IFS HRES vs Analysis - 10 days
ERA5-Forecasts vs ERA5 - 3 days
ERA5-Forecasts vs ERA5 - 5 days
ERA5-Forecasts vs ERA5 - 10 days
Keisler (2022) vs ERA5 - 3 days
Keisler (2022) vs ERA5 - 5 days
Keisler (2022) vs ERA5 - 10 days
Pangu-Weather vs ERA5 - 3 days
Pangu-Weather vs ERA5 - 5 days
GraphCast vs ERA5 - 3 days
GraphCast vs ERA5 - 5 days
GraphCast vs ERA5 - 10 days
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Q bias [g/kg]

Figure S10: Global mean 700hPa specific humidity bias for 3, 5 and 10 day lead times.

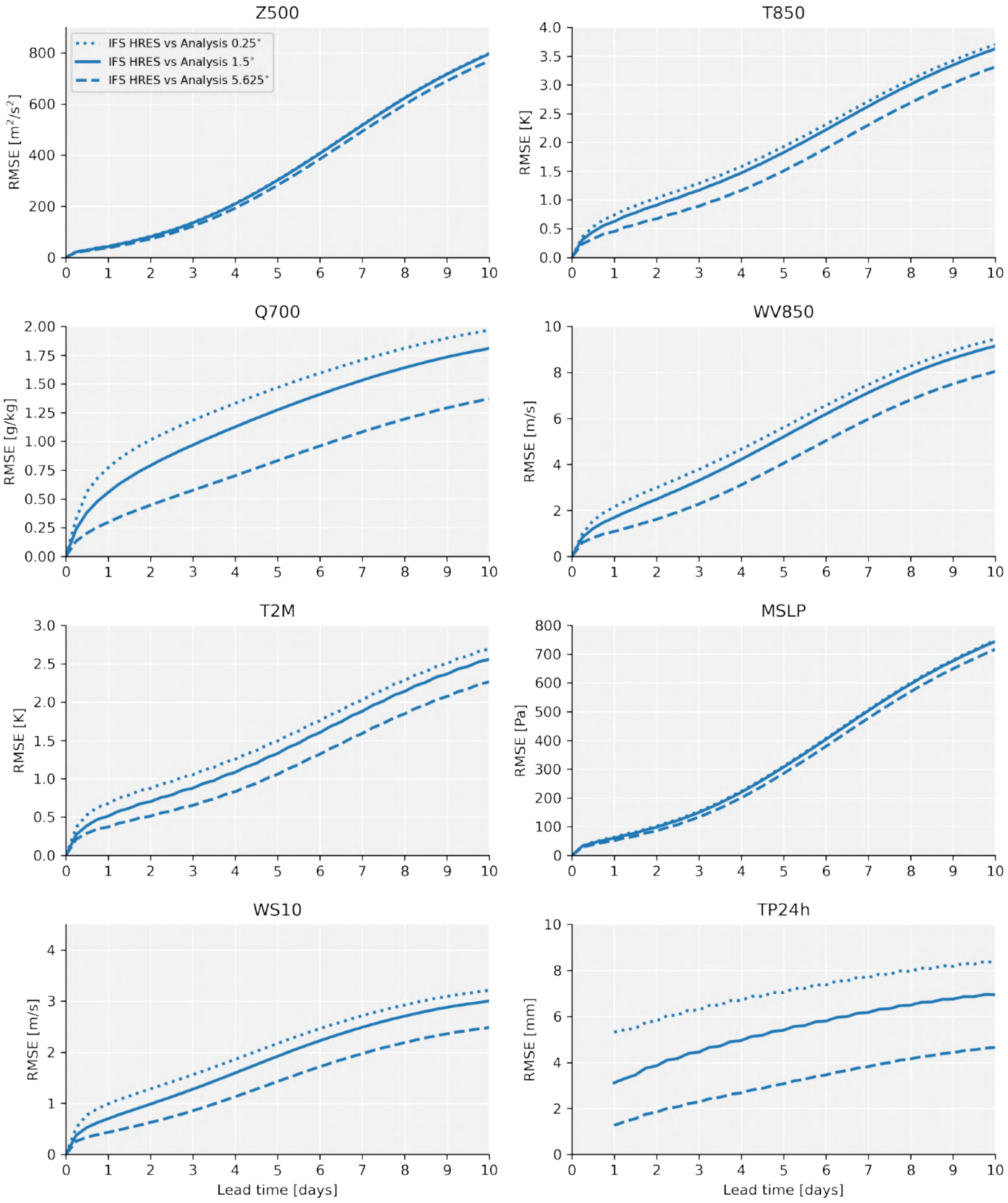

Figure S11: Comparison of RMSE scores for IFS HRES evaluated at different resolutions

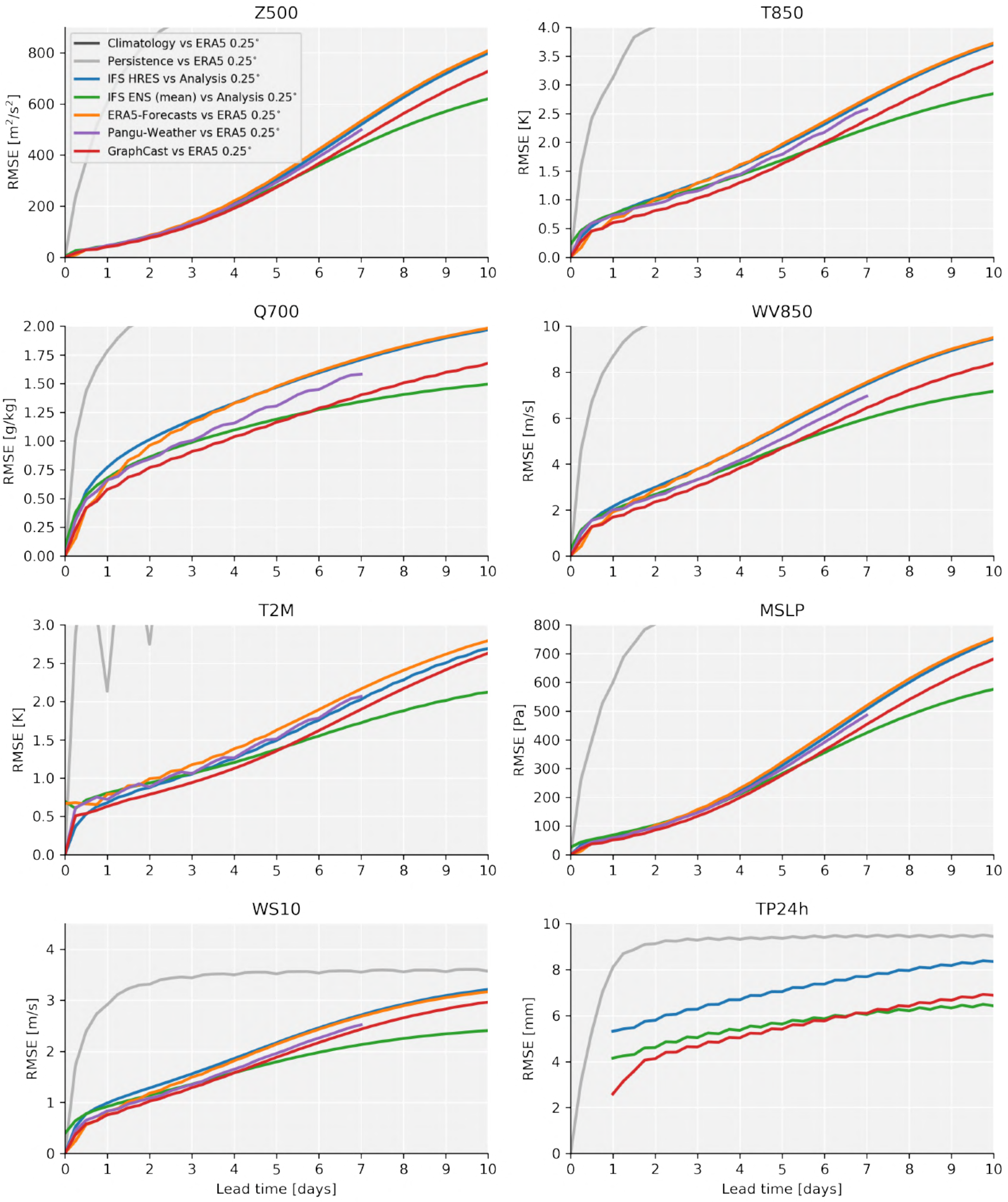

Figure S12: All models evaluated at 0.25° resolution.

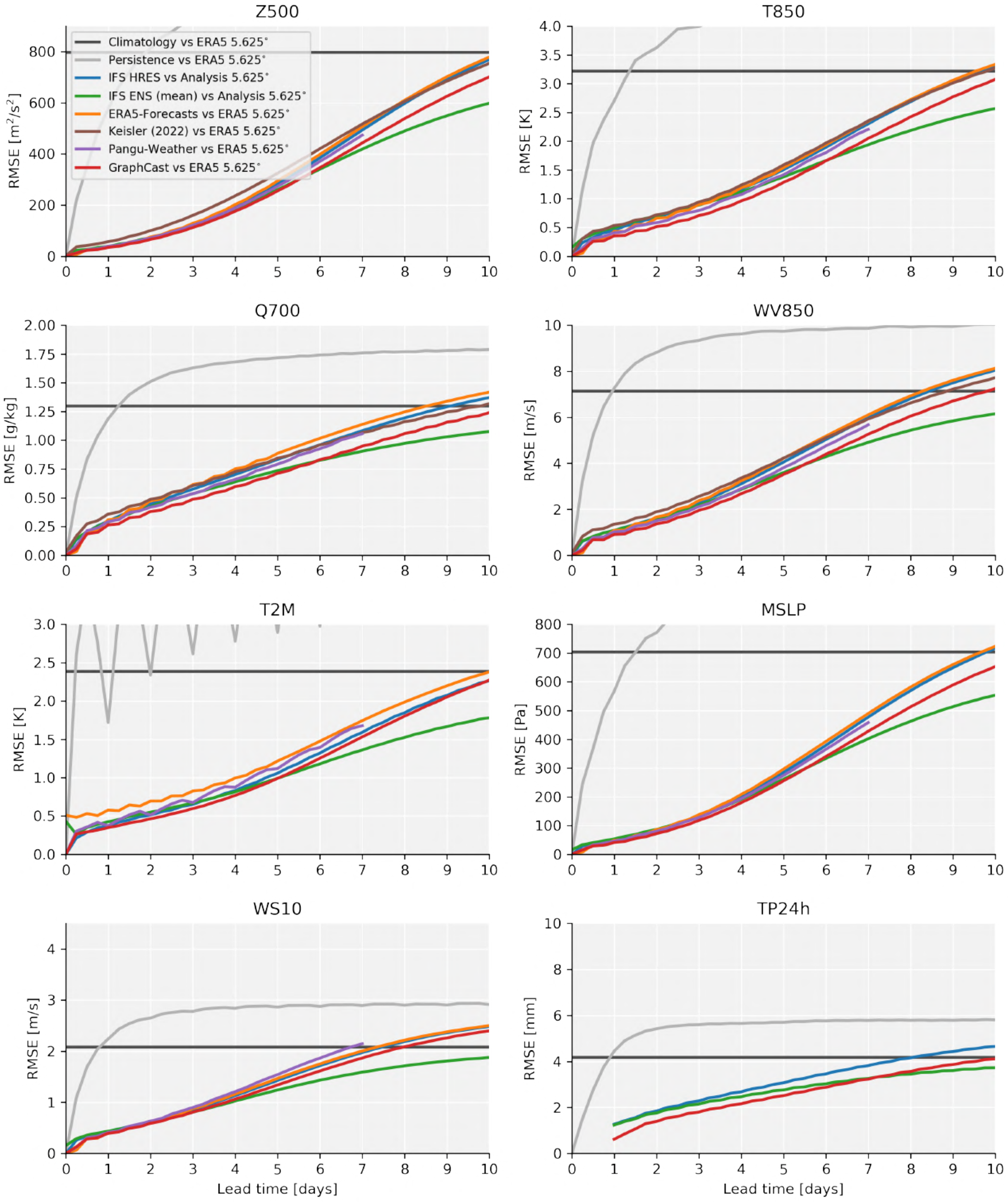

Figure S13: All models evaluated at 5.625° resolution.

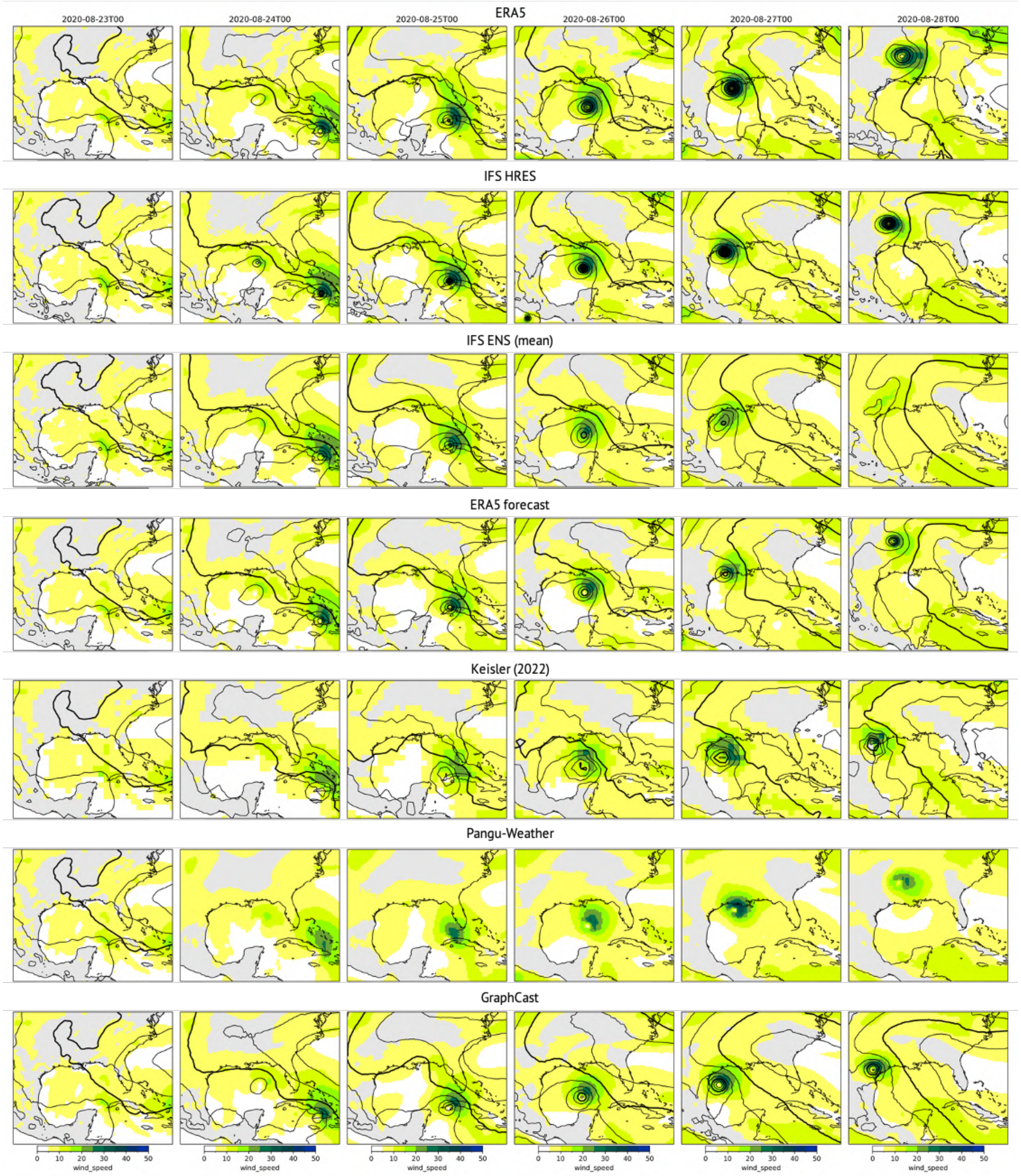

Figure S14: Case study: Hurricane Laura. Top row shows ERA5 ground truth. Other rows show forecasts initialized at 2020-08-23 00UTC. Lines show 850hPa geopotential contours (not available in our Pangu-Weather dataset). Shading shows 850hPa wind speed.