JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2021MS002502

Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models

Jonathan A. Weyn[1,3], Dale R. Durran[1], Rich Caruana[2], and Nathaniel Cresswell-Clay[1]

[1]Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA, [2]Microsoft Research, Redmond, WA, USA, [3]Microsoft, Redmond, WA, USA

Key Points:

- An ensemble forecast system is developed using convolution neural networks (CNNs) to generate data-driven global forecasts
- Only 3 s are required to compute a large 320-member ensemble of skillful 6-week sub-seasonal predictions
- Shorter lead time forecasts also show skill, including a single deterministic 4-day forecast for Hurricane Irma

Correspondence to:
J. A. Weyn,
jweyn@uw.edu

Citation:
Weyn, J. A., Durran, D. R., Caruana, R., & Cresswell-Clay, N. (2021). Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *Journal of Advances in Modeling Earth Systems*, *13*, e2021MS002502. https://doi.org/10.1029/2021MS002502

Received 9 FEB 2021
Accepted 19 JUN 2021

Abstract We present an ensemble prediction system using a Deep Learning Weather Prediction (DLWP) model that recursively predicts six key atmospheric variables with six-hour time resolution. This computationally efficient model uses convolutional neural networks (CNNs) on a cubed sphere grid to produce global forecasts. The trained model requires just three minutes on a single GPU to produce a 320-member set of six-week forecasts at 1.4° resolution. Ensemble spread is primarily produced by randomizing the CNN training process to create a set of 32 DLWP models with slightly different learned weights. Although our DLWP model does not forecast precipitation, it does forecast total column water vapor and gives a reasonable 4.5-day deterministic forecast of Hurricane Irma. In addition to simulating mid-latitude weather systems, it spontaneously generates tropical cyclones in a one-year free-running simulation. Averaged globally and over a two-year test set, the ensemble mean RMSE retains skill relative to climatology beyond two-weeks, with anomaly correlation coefficients remaining above 0.6 through six days. Our primary application is to subseasonal-to-seasonal (S2S) forecasting at lead times from two to six weeks. Current forecast systems have low skill in predicting one- or 2-week-average weather patterns at S2S time scales. The continuous ranked probability score (CRPS) and the ranked probability skill score (RPSS) show that the DLWP ensemble is only modestly inferior in performance to the European Center for Medium Range Weather Forecasts (ECMWF) S2S ensemble over land at lead times of 4 and 5–6 weeks. At shorter lead times, the ECMWF ensemble performs better than DLWP.

Plain Language Summary The world's leading weather forecasting institutions currently rely on computationally expensive weather models running on massive supercomputers. In order to have predictive skill for forecasts two to six weeks in the future, large ensembles of many nearly identical runs of these models are required, but the computational resources needed for these ensembles scales with the number of forecasts run. Since the resources needed rapidly approaches modern-day computing limits, we explore the possibility of using computationally cheap weather models based on machine learning algorithms which learn to reproduce the evolution of weather. Our machine-learning model is capable of running 320 forecasts in three minutes on a single workstation, while the state-of-the-art model from the European Center for Medium-Range Weather Forecasts (ECMWF) utilizes supercomputers to run 50 forecasts. Our ensemble weather model produces realistic forecasts of weather events such as Hurricane Irma in 2017 and is even capable of nearly matching the performance of the ECMWF ensemble for forecasts of temperature four to six weeks in the future.

1. Introduction

Weather forecasting relies heavily on data assimilation to estimate the current state of the atmosphere and on numerical weather prediction (NWP) to approximate its subsequent evolution. The skill of such deterministic weather forecasts is typically limited to about two weeks by the chaotic growth of small initial errors and inaccuracies in our approximate models of the atmosphere. On much longer, multi-month time scales, the coupling of the atmosphere with slowly evolving ocean-land forcing allows skillful seasonal forecasts of monthly or seasonally averaged conditions. Between these two extremes, the production of skillful one- or two-week averaged forecasts at lead times ranging roughly between two weeks and two months (the subseasonal-to-seasonal or S2S time frame) has proven particularly challenging; yet there are many societal sectors that would greatly benefit from improved S2S forecasts (White et al., 2017). Several major operational centers have developed NWP-based ensemble systems focused on improving S2S forecasting (Vitart et al., 2017).

In 1992, the European Center for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction (NCEP) began issuing ensemble weather forecasts. They were soon followed by the other major weather prediction centers across the world. Ensemble forecasts strive to provide a set of equally likely forecast realizations spanning the range of possible future atmospheric states. Early evidence for the economic value of probabilistic forecasts derived from the ECMWF ensemble relative to a single deterministic forecast was provided by Richardson (2000). Ensemble forecasts are now recognized as essential to represent the probabilistic nature of weather forecasting and to break through the intrinsic limits to predictability of the atmosphere (Palmer, 2018).

Ensembles are particularly appropriate as one looks beyond lead times where deterministic forecasts lose all skill relative to climatology. On S2S lead times, ensemble-mean and ensemble-based probabilistic forecasts have shown modest skill relative to climatology (Monhart et al., 2018; Vitart, 2004, 2014; Weigel et al., 2008). Computational resources do, however, impose a significant limitation on efforts to create S2S forecasts with NWP ensembles. As of 2016, 11 forecast centers were contributing S2S forecasts to the S2S database (Vitart et al., 2017), and the ensembles from three of these centers consisted of just four members. The largest S2S ensemble, with 51 members providing forecasts out to 46 days, is generated by ECMWF's Integrated Forecast System (IFS) using a nontrivial fraction of the time available on one of the world's most powerful computer systems (Bauer et al., 2015). Yet there is increasing evidence that the number of ensemble members for S2S forecasts should be higher, perhaps in the 100–200 range (Buizza, 2019). Large ensemble sizes are also helpful for assessing the likelihood of events in the tails of probabilistic forecast distributions (Leutbecher, 2018), and such extreme events are often the most impactful.

Machine learning provides one potential avenue to develop S2S forecasts systems with significantly lower computational costs. Recognizing that there are other successful machine-learning approaches to S2S forecasting (Hwang et al., 2019; Mayer & Barnes, 2020), here our focus will be on the development of a data-driven deep-learning weather prediction (DLWP) model that can be iteratively stepped forward, like traditional NWP models, to simulate atmospheric states at arbitrarily long lead times. In one of the first attempts to use ML to create such a model, Dueben and Bauer (2018) trained neural networks (NNs) on several years of reanalysis data to predict 500 hPa geopotential height (Z_{500}) on the globe at 6° resolution, demonstrating the ability to produce ML weather forecasts that have at least modest forecast skill. Using advanced convolutional neural networks (CNNs), Scher and Messori (2019) trained algorithms on simulations from a simplified GCM that significantly outperformed baseline metrics and effectively captured the simplified-GCM dynamics at spherical harmonic resolutions of T21 and T42 (roughly 5.6°and 2.8°). Training only on reanalysis data, Weyn et al. (2019) used CNNs to generate forecasts for northern mid-latitude Z_{500} and 300–700-hPa thickness ($\tau_{300-700}$) on a 2.5° latitude-longitude grid that showed skill relative to climatology and persistence through five days.

Recently we extended our DLWP model to the full globe using a volume-conservative mapping to project global data from latitude-longitude grids onto a cubed sphere and improved the CNN architecture operating on the cube faces (Weyn et al., 2020, hereafter WDC20). In addition to Z_{500} and $\tau_{300-700}$, our improved model forecasts two additional surface fields, 1,000 hPa height (Z_{1000}) and 2-m temperature (T_2), and uses three externally specified 2D fields: a land-sea mask, topographic height, and top of the atmosphere radiation. This new 1.9° resolution model showed skill relative to climatology and persistence through seven days. Moreover, it could be stepped forward repeatedly from a single initialization for at least one year, and while doing so, captured the seasonal cycle with reasonable accuracy.

In the following we further improve our DLWP model by adding two more 2D prognostic fields and increasing the spatial resolution to 1.4°. Large 320-member ensembles generated using the improved model are used to provide S2S forecasts through a six-week lead time. These forecasts are verified against ERA5 data and compared to operational ECMWF S2S products.

The remainder of this paper is organized as follows. Section 2 describes the improvements to our previous DLWP model, while section 3 discusses the incorporation of that model in our ensemble forecast system. The behavior of that ensemble is assessed at short deterministic lead times in Section 4, and at longer S2S lead times in Section 5. Section 6 contains the conclusions.

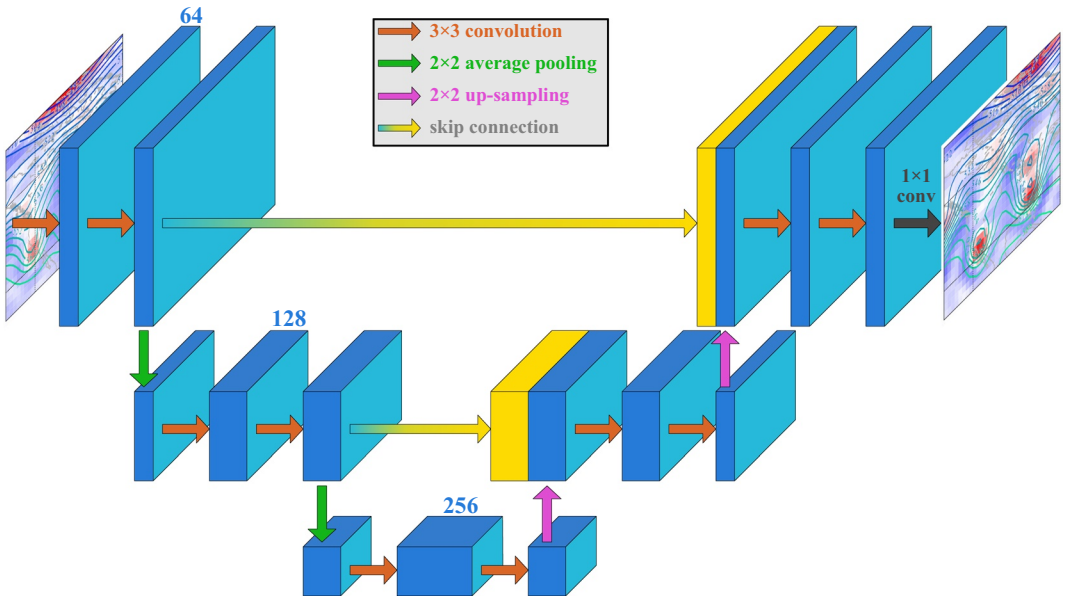

Figure 1. Schematic illustrating the architecture of our DLWP CNN based on the U-Net architecture. Each red arrow represents a 2-D convolution operating on each cubed-sphere face. Green and purple arrows indicate average-pooling and upsampling operations, respectively. The blue-to-yellow lines represent skip connections, whereby the blue state is copied exactly to the yellow state vector and concatenated to the new blue state vector along the channels dimension. The final black arrow is a 1 1 convolution. The blue numbers indicate the number of convolutional filters (channels) at each stage of the network (channel width is to scale).

2. The DLWP Model

The basic model is very similar to that described in detail in WDC20, in which four forecast fields, geopotential height at 1,000 hPa (Z_{1000}) and at 500 hPa (Z_{500}), 300–700 hPa thickness ($\tau_{300-700}$), and 2-m temperature (T_2), are mapped to a cubed sphere. Three known fields are also provided: top-of-atmosphere radiation, topographic height, and a land-sea mask. Convolutional neural networks (CNN) were trained using the same 3×3 set of horizontal spatial filters on all four equatorial faces on the cube (WDC20, Figure 1). A different set of filters was applied to the two polar faces. A U-Net architecture (Ronneberger et al., 2015) with skip connections is employed to capture multi-scale processes via average pooling and corresponding up-sampling. The skip connections across each level of spatial refinement ensure high-resolution information is preserved. The activation functions are leaky ReLU functions capped at a scaled value of 10. The model is recursively stepped forward with 12-h time steps, such that a single step maps the fields at two time levels $t_0 - 6$ and t_0 hr to forecast fields at $t_0 + 6$ and $t_0 + 12$ h. The model is trained to minimize the mean squared error (MSE) of all forecast fields over two steps, or equivalently over a 24-h period with 6-h temporal resolution. This iterative approach follows the time-stepping strategy in NWP; once our model is trained, it can provide forecasts at any lead time that is a multiple of six hours. This approach may contrasted with models trained to forecast only for specific lead times, such as 6 hours, 1, 3, and 5 days in Rasp and Thuerey (2021).

Here we extend the WDC20 model by adding two more forecast fields: temperature at 850 hPa (T_{850}), which is strongly modulated by large-scale weather patterns while exhibiting less sensitivity to diurnal heating and surface-layer processes than T_2, and total column water vapor (TCWV). TCWV is the vertically integrated total gas-phase water above each grid cell; its inclusion is a step toward characterizing tropical convective systems including tropical cyclones and the Madden-Julian oscillation (MJO). The MJO is believed to be an important source of forecast skill on S2S time scales and has been characterized as a "moisture mode" (Adames & Kim, 2016).

The resolution was also modestly increased from 48×48 to 64×64 grid cells on each face of the cube, yielding an effective resolution of approximately 1.4° in latitude and longitude at the equator. ERA5 data at a gridded resolution of 1° in latitude and longitude were remapped with the Tempest-Remap package (Ullrich

Table 1
CNN Architecture for DLWP as a Sequence of Operations on Layers

Layer	Filters	Filter size	Output shape[a]	Trainable params[b]
input			(6, 64, 64, $vt + c$)	
Conv2D–CubeSphere	64	3 × 3	(6, 64, 64, 64)	18,560
Conv2D–CubeSphere (1)	64	3 × 3	(6, 64, 64, 64)	73,856
AveragePooling2D		2 × 2	(6, 32, 32, 64)	
Conv2D–CubeSphere	128	3 × 3	(6, 32, 32, 128)	147,712
Conv2D–CubeSphere (2)	128	3 × 3	(6, 32, 32, 128)	295,168
AveragePooling2D		2 × 2	(6, 16, 16, 128)	
Conv2D–CubeSphere	256	3 × 3	(6, 16, 16, 256)	590,336
Conv2D–CubeSphere	128	3 × 3	(6, 16, 16, 128)	590,080
UpSampling2D		2 × 2	(6, 32, 32, 128)	
Concatenate (2)			(6, 32, 32, 256)	
Conv2D–CubeSphere	128	3 × 3	(6, 32, 32, 128)	590,080
Conv2D–CubeSphere	64	3 × 3	(6, 32, 32, 64)	147,584
UpSampling2D		2 × 2	(6, 64, 64, 64)	
Concatenate (1)			(6, 64, 64, 128)	
Conv2D–CubeSphere	64	3 × 3	(6, 64, 64, 64)	147,584
Conv2D–CubeSphere	64	3 × 3	(6, 64, 64, 64)	73,856
Conv2D–CubeSphere	vt	1 × 1	(6, 64, 64, vt)	1,560

Note. The parameter v represents the number of input fields, t represents the number of input time steps, and c represents the number of auxiliary prescribed inputs (here top-of-atmosphere radiation at t times, land-sea mask and topographic height). The layer names (except for the suffix "CubeSphere") correspond to the names in the Keras library. "Concatenate" appends the state in parentheses, numbered earlier, to the output of the previous layer.

[a]Output shape is (face, y, x, channels). [b]Number of learned parameters for $t = 2$, $v = 6$, $c = 4$. Total is 2,676,376.

& Taylor, 2015; Ullrich et al., 2016) for training, validation and testing. ERA5 data from years 1979–2012 were used for training, 2013–2016 for validation, and 2017–2018 for the final test set.

The WDC20 convolutional neural network architecture continued to perform quite well despite the changes to the model input and target data. We were able to improve the model further by doubling the number of filters used in each convolutional layer. This increased the number of filters in the first layer from 32 to 64. The architecture of our revised U-Net is diagrammed in Figure 1 and tabulated in Table 1. The increases in the number of filters in each layer and number of forecast fields increased the total number of trainable parameters relative to that in WDC20 by a factor of about 4, to 2.7 million. Nevertheless, as a result of additional code optimization, the model still trains in 6–8 days on a single Nvidia Tesla V100 GPU.

3. Designing an Ensemble of DLWP Models

The basic ensemble design follows the typical practice used in operational NWP forecasting by including ensemble members with both perturbed initial conditions (ICs) and variations in the model's representation of the atmosphere—the latter being incorporated in NWP ensembles either through the use of several different "physics" parameterization packages, through a suite of different parameter values with a fixed set of packages, or through the incorporation of stochastic physics. The perturbed initial conditions and our approach to varying the model representation of the atmosphere are discussed below.

3.1. Initial Condition Uncertainty

The ERA5 data set includes 10 perturbed ensemble members generated by ensemble data assimilation with 4DVAR (Isaksen et al., 2010) to help with uncertainty estimation, and we use these as a convenient set of perturbed ICs for construction of our DLWP ensemble. Unfortunately, this set of ICs is non-optimal, because, unlike the operational ECMWF ensemble (Palmer, 2018), singular vectors were not used to select the most rapidly growing initial perturbations. Moreover, the ERA5 ensemble itself is moderately under-dispersive (https://confluence.ecmwf.int/display/CKB/ERA5%3A+uncertainty+estimation).

Figure 2 shows the globally averaged RMSE of the ensemble mean T_{850} plotted as a function of forecast lead time for four DLWP ensemble strategies. Also plotted are RMSE reference curves for climatology and persistence. The solid blue curve shows the RMSE of a DLWP forecast generated using the 10 perturbed members of the ERA5 data set to create a 10-member IC ensemble. Since the ERA5 IC perturbations do not project strongly on to the most rapidly growing modes, the ensemble spread (computed as the square root of the average over all forecasts of the ensemble variance) actually decreases over the first 36 h of forecast lead time (dashed blue line in Figure 2). In a conventional NWP model, such a reduction in ensemble spread, which occurs primarily on small spatial scales, can arise from a combination of numerical dissipation and the dispersion of inertia-gravity waves, and an analogous behavior is present in the DLWP model.

Another serious problem with the IC ensemble is that the spread is much smaller than the RMSE of the ensemble mean. In an ideal ensemble, the joint distribution of the ensemble members would be unchanged if the verifying observations were substituted for any one of the individual ensemble members, and under that assumption the spread and RMSE curves should coincide (Fortin et al., 2014). In an effort to better match the ensemble spread to the RMSE, we ran a sensitivity test using a 10-member "IC × 2" ensemble, in which the difference in the IC perturbations from the control ERA5 data was doubled. The spread for

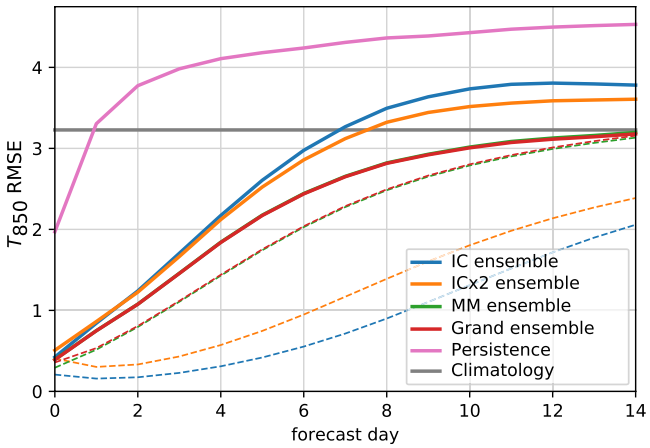

Figure 2. RMSE of T_{850} (K) as a function of forecast lead time for DLWP ensembles (solid lines) and corresponding ensemble spread (dashed): IC (blue), IC × 2 (orange), multi-model (MM; green) and the grand ensemble (red). Curves for the MM and grand ensemble are almost identical. Also shown are the RMSE for persistence (pink) and climatology (gray) benchmarks.

the IC × 2 ensemble is modestly improved, and over the period 5–14 days the RMSE is modestly reduced (orange curves) relative to the original IC ensemble. Note that the reduction in initial ensemble spread over the first 36 h is more rapid in the IC × 2 ensemble, providing further evidence that the variations in these initial condition do not strongly project on the structure of the most rapidly growing perturbations. Because this ad-hoc methodology did not significantly improve the spread-skill ratio of the IC ensemble, we retained the original perturbations for the grand ensemble presented through the rest of this paper.

3.2. Uncertainty in the Representation of the Atmosphere

The learnable weights for convolutions in the CNN are initialized as small random values, and we can exploit this randomness by repeatedly retraining with different initial seeds to produce a family of DLWP models with slightly different final weights. The models in this family are capable of making approximately equally skillful forecasts, but with enough statistical independence to produce a good ensemble. Scher and Messori (2020) pursued a similar strategy, training and re-training models to both emulate a simple GCM and to forecast the real atmosphere with data from ERA5.

As part of its holistic estimation of the next atmospheric state, our CNN-based DLWP model effectively captures physical processes that are parameterized in operational NWP models. For example, as noted in WDC20 (see also Figures 6c and 6d), our DLWP model does an excellent job of forecasting 2-m temperatures, including their diurnal cycle, without using any explicit parameterization of boundary layer processes, and without including most of the meteorological fields that would be used in such parameterizations. We will refer to the ensembles in which the CNN filter coefficients are randomly perturbed during training as "multi-model" (MM) ensembles, because forming an ensemble in this manner is analogous to forming an ensemble of GCMs by specifying slightly different constants in their physical parameterizations.

Rather than completely retrain each member of our MM ensemble from new random seeds, we gained efficiency by using intermediate results produced during the training process. Our DLWP model is trained using the adaptive learning scheme Adam (Kingma & Ba, 2014). Figure 3 shows the learning curve for our loss function (mean squared error, see WDC20) as a function of the training epoch number for a training cycle representative of that for one of our DLWP ensemble members. As expected, the error on the training set decreases smoothly, while the error on the validation set oscillates much more, strongly suggesting the learned weights in the model undergo nontrivial changes over each training epoch. The variations induced by these changes in the weights turn out to be sufficient to provide many useful members in a MM ensemble, and we exploited these variations as follows.

After at least 100 training epochs, we selected multiple potential ensemble members from a single training cycle by checkpointing every 10 epochs and saving the model's weights at each checkpoint. Model training was ended when the validation loss did not improve for 50 consecutive epochs. Using the checkpointed weights, we tested each resulting model's utility for S2S forecasts by evaluating its average global T_{850} anomaly correlation coefficient (ACC) score at 4-week lead time from twice-a-week forecasts over the full four-year validation set. The 4-week T_{850} ACC scores among the top performers for each training cycle were nearly indistinguishable from each other and from those of other training cycles, and there was nearly as much ensemble spread between the various checkpoints of one training cycle as there was between models generated by separate training cycles. (Although the final trained model had the lowest validation set loss during training, the training loss function is based solely on the RMSE over the first 24 h of forecast lead time; it does not necessarily reflect on the performance of a 4-week forecast.)

Our MM ensemble used a total of 32 slightly different DLWP models generated by eight training cycles with four members drawn from the checkpoints of each cycle. The members selected from each cycle had the

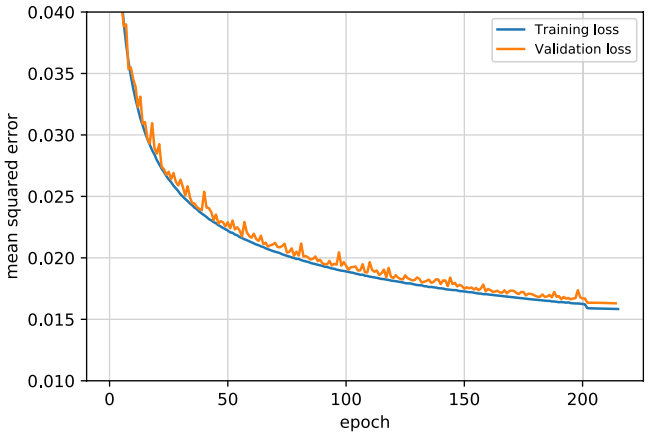

Figure 3. Example of the CNN learning curve for a representative DLWP model. The loss, which is the mean-squared-error over all scaled target outputs of the model, evaluated on the training (validation) set is shown as a function of training epoch number in blue (orange). The optimizer was switched from Adam to a standard stochastic gradient descent (SGD) optimizer after 200 epochs, producing an abrupt small decrease in the loss function on both data sets and more uniform values on the validation set.

best 4-week T_{850} ACC scores, although three of the 32 members selected in this fashion required further refinement because in a few individual forecasts they produced fields with unrealistic structures even though their numerical values remained bounded within reasonable limits. Those three members were further trained for two more epochs using a re-initialized stochastic gradient descent (SGD) optimizer, which consistently fixed the issue with nonphysical solutions. The SGD cycles also lowered the training and validation loss slightly, as exemplified over the last 15 epochs of the learning curve in Figure 3.

The RMSE and spread of the MM ensemble is plotted in Figure 2. It clearly outperforms the IC and IC × 2 ensembles, having lower RMSE and a much better RMSE-spread relationship. The spread is roughly 80% of RMSE at a forecast lead time of 8 d and approaches 95% by 14 d. At 14 d, the RMSE of the MM ensemble remains slightly better than climatology, whereas the RMSE for the IC and IC × 2 ensembles begins to exceed climatology between 7 and 8 d. Scher and Messori (2020) compared the performance of initial-condition and retrained (i.e., MM) neural network ensembles. For the 500-hPa forecasts using the less accurate DLWP model of Weyn et al. (2019), they found a relationship between the RMSE and spread in their MM ensemble roughly similar to that in Figure 2, although in their case the spread matched the RMSE at the end of their 5-day forecast and the RMSE for 500-hPa height was roughly twice as large as that for our current model at the same 5-day forecast lead time (compare Figure 6a with their Figure 4a). In contrast to our IC ensembles, their ICs were generated using singular vectors and produced only slightly less spread than their MM approach. Finally, as in our results, the RMSE of their IC ensemble mean was higher than the RMSE for their MM ensemble mean.

Scher and Messori (2020) did not report results for a grand ensemble consisting of both IC and MM perturbations, but given the superior ensemble spread they obtained using singular vectors, such a grand ensemble might have performed substantially better than either of the individual IC or MM ensembles. In our case, a 320-member grand ensemble constructed by applying the suite of 32 DLWP models with slightly different weights to each of the 10 IC perturbations performs only very slightly better than the MM ensemble alone at 14-d forecast lead times: the RMSE and spread curves (red) for the grand ensemble almost perfectly overlap those for the MM ensemble (green) in Figure 2. Nevertheless, after bias correction (see section 3.4) the 320-member grand ensemble does perform better than the MM ensemble at longer S2S forecast lead times.

3.3. The Control Member

When considering the effectiveness of ensemble forecasts, it is useful to compare the ensemble mean to a single control member. For example, in comparison to the other ensemble members, the ECMWF control forecast is run at higher horizontal (9 km) and vertical (137 levels) resolution, and without perturbations to the initial conditions. This control forecast might nominally be expected to perform better than a typical ensemble member that must be run at lower resolution because of computational constraints. Our control forecast is trained to better minimize the loss function using the Adam optimizer and learning rates that decrease as the weights approach their optimum values. Specifically, the learning rate starts at 10^{-3}, but once the validation-set loss does not decrease for 20 epochs, the learning rate of the optimizer is reduced by a factor of 5. This continues (up to a minimum learning rate of 10^{-6}) until the same criterion of no reduction in validation loss after 50 epochs is met. The result is a model with weights that better minimize the loss function and produces good forecasts, although in contrast to the ECMWF high-resolution control, the model used for our control forecast does not have significant advantages relative to the other ensemble members. The control forecast is initialized with the control ERA5 reanalysis data.

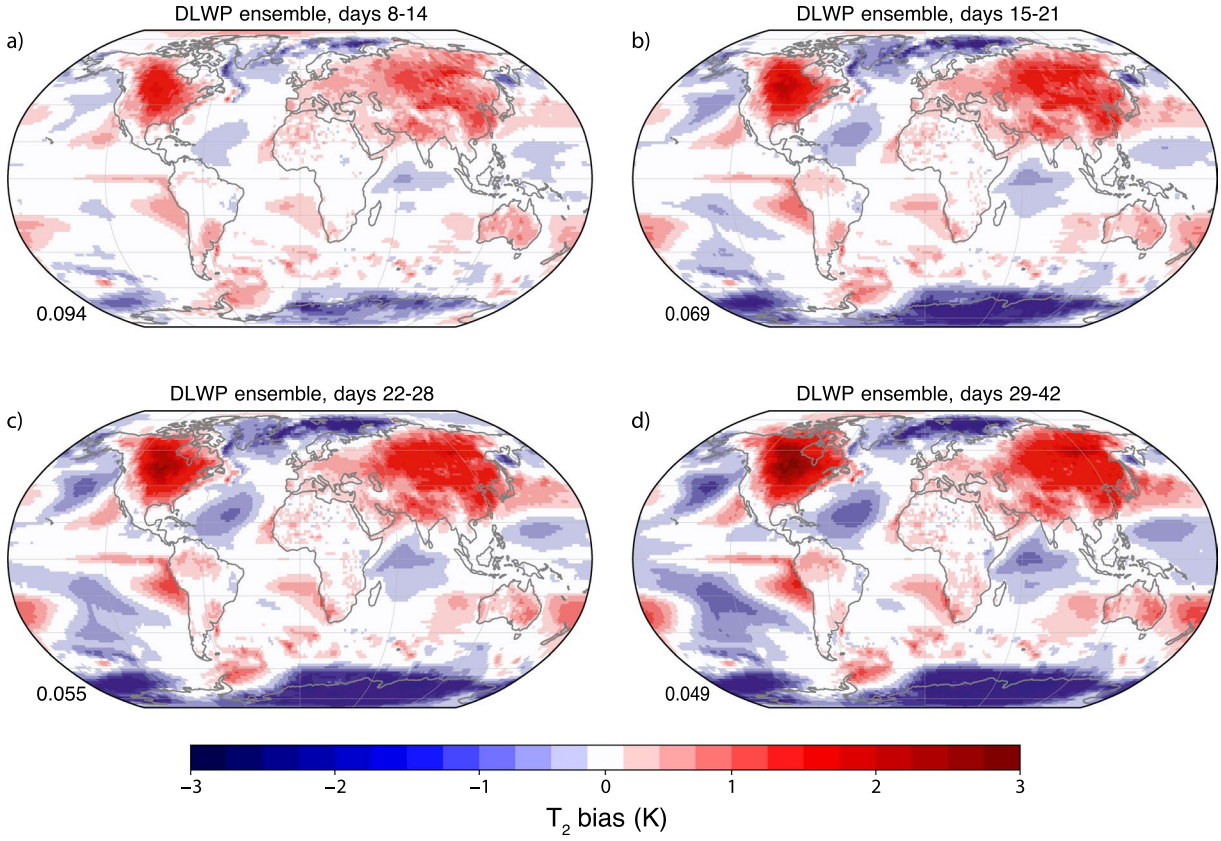

Figure 4. Bias in 2-m temperature (ensemble-mean − observation) averaged over 25 years of re-forecasts from all members of the DLWP MM ensemble: one-week averages for forecast lead times of (a) 2 weeks, (b) 3 weeks, (c) 4 weeks, and (d) a 2-week average for the week 5–6 forecast. Globally-averaged bias is shown as the number in the lower-left of each panel.

3.4. Correcting Model Bias

Unlike global climate models, NWP models are designed to make accurate predictions over relatively short forecast lead times without worrying about certain physical constraints, such as global radiative balance, that would be necessary for long-term climate simulations. As a consequence, NWP models are typically subject to systematic drift in long-term (including sub-seasonal) forecasts. For example, previous versions of the ECMWF S2S ensemble have been shown to develop pronounced spatially dependent patterns of mean model drift on time scales of 1–4 weeks (Vitart, 2004; Weigel et al., 2008). To compensate for this model bias, all of the major weather prediction centers produce reforecasts, or hindcasts, using their S2S ensemble prediction systems (Vitart et al., 2017), and use these reforecasts to calibrate the operational forecast products. As an example, Vitart (2004) computed the bias for a given calendar date by performing hindcasts using the full ensemble initialized on the same calendar date in each of 12 previous years, and removed this bias from the forecasts in a post-processing step.

Using a strategy similar to that in Vitart (2004), we bias corrected each field by first computing ensemble reforecasts twice weekly for each DLWP MM ensemble member (and the control) on the same set of calendar dates spanning the years 1991–2015, which includes the training set and most of the validation set, but does not bleed into the 2017–2018 test set. Then, for all reforecast dates, each member's spatially varying bias is calculated as the average of its bias on that date over the 25 year period. The bias for a specific forecast date in the test set is taken as the average reforecast bias over all available calendar dates spanning the 28-day interval centered on that forecast date, and for each field this bias is subtracted from the forecast produced by each of the corresponding ensemble members.

Bias correction was not used to obtain the results shown in Figure 2, but it is used in all our subsequent analyses. Figure 4 shows spatial patterns of the annual-average model bias in 2-m temperature for the

Table 2
Comparison of Key Attributes of Our DLWP Ensemble and Those of the State-of-the-Art ECMWF Ensemble for Extended-Range Forecasting

	DLWP	ECMWF
Atmospheric fields	6 2-D variables	9 prognostic 3-D variables; 91 vertical levels
Horizontal resolution	150 km	18 km (36 km after day 15)
Atmospheric physics	3 prescribed inputs	Many physical parameterizations
Coupled models	None	Ocean, wave, and sea ice models
Initial condition perturbations	10 (ERA5 uncertainty)	50 (SVD/4DVAR)
Model perturbations	Perturbed CNN weights	Stochastic physics
Ensemble members	320 (+control)	50 (+control)

32-member DLWP MM ensemble at forecast lead times up to 6 weeks. The magnitude of the bias is roughly comparable to that shown for 2-m temperature in Weigel et al. (2008). Warm biases are present over the northern hemisphere land masses, along with a cold bias over Antarctica. There are also warm biases in subtropical regions commonly dominated by marine stratocumulus clouds off the Pacific coasts of North and South America. These biases gradually amplify as the forecast lead time increases, although the globally averaged spatial-mean bias (noted in each panel) decreases at longer lead times. The tendency of increasing local biases to better cancel in the global mean at longer lead times is interesting and perhaps surprising because the model is only trained to minimize T_2 errors over the first 24 h of the forecast—no global energy-balance constraints are imposed.

Bias correction has a positive impact on the control forecast and on the IC, MM, and the grand ensembles. Although the RSME and spread of the MM and grand ensembles are almost identical over the first 14 days, at longer lead times, and particularly after bias-correction, the grand ensemble is clearly superior to the MM ensemble (not shown). The performance of the grand ensemble will, therefore, be our focus throughout the remainder of this paper.

3.5. ECMWF Ensemble Benchmark

In addition to the persistence and climatology benchmarks, which serve as a baselines that must be exceeded by any skillful forecast, we will also compare our results against the state-of-the-art ECMWF 50 member S2S ensemble and a higher resolution ECMWF control simulation (Vitart et al., 2017). Errors are computed with respect to ERA5 data that is downloaded at 1° resolution, transformed onto our cube-sphere grid, and then transformed back to a 1.5 × 1.5 latitude-longitude grid. Our DLWP forecasts are transformed to the same 1.5 × 1.5° grid for the computation of all forecast metrics. The archived ECMWF S2S forecasts, available on a 1.5 × 1.5° grid, are first transformed to the cube sphere and then back to the 1.5 × 1.5° analysis grid because this procedure removed discrepancies in model terrain thereby improving the ECMWF error metrics for T_2 when evaluated against the same validation data as the DLWP ensemble. Bias correction was also performed on the ECMWF S2S control and ensemble forecasts on the 1.5 × 1.5° grid, with the methodology following that of the operational ECMWF forecasts. This correction is very similar to the bias correction applied to our DLWP model, but with a few differences: the last 20 years of reforecasts are used instead of a fixed period of 25 years; 10 ensemble members with perturbed IC and physics are run for each reforecast; and only the forecasts for dates within one week, instead of 28 days, of the target operational forecast issue date are used.

3.6. Summary

The following summarizes the construction of the DLWP grand ensemble.

1. Eight distinct training cycles of the DLWP CNN were produced with different random seeds as a first step in generating a multi-model ensemble with 32 members.

2. Four checkpoints during each of the eight training cycles were selected based on T_{850} ACC skill as individual MM ensemble models.
3. The model associated with each checkpoint was run on the validation data set to produce 416 four-week forecasts, which were then manually inspected for forecast quality. Any model (a total of three) that displayed irregularities was further trained with an SGD optimizer. The collection of models given by these checkpoints formed the 32-member MM ensemble.
4. Each of the 32 MM models was run with each of the 10 initial conditions (ICs) given by the perturbed reanalyses in the ERA5 product to yield the 320-member grand ensemble.
5. A single control DLWP model was trained slightly differently, by periodically reducing the Adam optimizer learning rate.
6. The mean model bias for reforecasts in the period 1991–2015 was computed for the control and each member of the MM ensemble. That bias was removed from all the 2017–2018 test-set forecasts.

Finally, a tabular comparison between our DLWP ensemble and the current state-of-the-art ECMWF ensemble is provided in Table 2. In most regards, the ECMWF ensemble is superior, with higher resolution, coupled ocean and wave models, complete physics, more atmospheric variables, and better initial condition perturbations. However, our DLWP ensemble consists of far more ensemble members, with 320 plus control, compared to only 50 plus control for the ECMWF ensemble.

4. Skill of the DLWP at Short Lead Times

Before discussing the performance of the DLWP ensemble on S2S time scales, we briefly assess the qualitative skill of the model in a short deterministic forecast. We then assess the quantitative skill of the DLWP control and grand ensemble in two-week forecasts relative to the state-of-the-art 50-member ECMWF S2S ensemble and the standard benchmarks of climatology and persistence.

Figure 5 compares a 4.5-day global forecast of Z_{1000} and Z_{500} with the verifying analysis for 12:00 UTC, September 11, 2017, when Hurricane Irma was located in the southeastern US. Irma is farther north and weaker in the DLWP forecast, but still reasonably well-represented for a 4.5-day forecast of such a small-scale disturbance. Other features, such as the 500-hPa cutoff low west of California and the pair of shortwaves in the Gulf of Alaska and west of Hudson Bay are also reasonably represented. On the other hand, the 500-hPa cutoff low over Novia Scotia is much weaker and farther west than in the verification, and the DLWP forecast does not develop the associated surface cyclone. The closed surface low in the DLWP forecast over the Dominican Republic is the model's forecast for Hurricane Jose, which is stronger and farther south than in the verification. While neither perfect nor the equal of a state-of-the-art NWP operational forecast, on the balance the DLWP forecast is arguably still impressive given that it is computed at 1.4° resolution using just 6 prognostic variables, each defined on a single spherical shell.

Turning to the quantitative verification of the first two weeks of forecast lead time, our cases are chosen to match the available forecast initialization times from the operational S2S forecast runs at ECMWF. The DLWP model is therefore tested on 208 forecasts initialized twice weekly starting at 00 UTC January 2, 2017, followed by the fifth, ninth and twelfth of January, and so on, through the end of 2018.

RMSE scores for Z_{500} are compared in Figure 6a. The DLWP grand ensemble remains superior to climatology through 14 days, though unsurprisingly, its error exceeds that of the ECMWF S2S ensemble. Both the DLWP and ECMWF control forecasts perform worse than their respective ensembles, and the control-to-ensemble improvement is qualitatively similar for both systems. At lead times beyond 9 days, the DWLP ensemble performs better than the ECMWF control. Similarly qualitative behaviors are apparent for the Z_{500} anomaly correlation coefficient (ACC) scores shown in Figure 6b, although the ACC score for the ECMWF control remains superior to that of the DLWP ensemble until almost day 13. The persistence forecast is grossly inferior.

Turning now to the daily averaged surface temperature field, which has a far more complex structure than Z_{500}, Figures 6c and 6d show the performance of the DLWP models relative to the ECMWF system remains similar to that for Z_{500}. (Only the daily averaged T_2 field was available on the ECMWF archive. As shown in WDC20, the DLWP model does capture diurnal temperature variations.) The ECMWF S2S ensemble

gives the best results, with its RMSE, together with that of the DLWP grand ensemble, remaining below the climatological benchmark through 14 days. Note that the ECMWF product uses slightly different initial conditions, which explains the substantial error in the ECMWF RMSE at early lead times. The ACC of the DLWP ensemble again becomes superior to the ECMWF control at long lead times (after 11 days).

Error metrics for a third field, T_{850}, are plotted in Figures 6e and 6f. This is a more difficult field to forecast in the sense that the RMSE for persistence exceeds climatology at shorter lead times than for Z_{500} or T_2, and the ACC score for persistence drops below 0.4 in just 2 days. Nevertheless, the RMSE for the DLWP and ECMWF ensembles again remains below climatology for 14 days, with the ECMWF ensemble performing the best. The errors in the DLWP ensemble also drop below those in the ECMWF control at earlier lead times than for the other fields: 7 days for RMSE and 10 days for ACC.

5. Extending the Forecasts to the S2S Range

5.1. Ensemble-Mean Anomaly Correlations

Globally—and temporally-averaged ACC scores for T_2 and T_{850} at forecast lead times of 3–4 and 5–6 weeks are compared in Figure 7 for the DLWP and ECMWF models, along with the persistence benchmark, for the 2017–2018 test set. Persistence forecasts are computed by averaging the observed anomalies from the 14 days prior to the forecast initialization date. At both lead times, scores are higher for T_2 than for T_{850}, reflecting greater memory in the system for near-surface temperatures than those aloft. One source of this memory is sea surface temperature, which is closely tied to T_2 over the oceans. Unlike the ECWMF S2S model, our DLWP model does not currently include coupling with the ocean, and as a consequence, our T_2 forecast over the ocean is essentially a proxy forecast for SST. Despite this sub-optimal treatment of the influence of SST, the DLWP ensemble is superior to both persistence and the ECMWF control for T_2 forecasts at both lead times. Unsurprisingly, the ECMWF ensemble gives the best results at both lead times, with an averaged ACC of roughly 0.5 for T_2 at 3–4 weeks. Turning to T_{850}, the performance of the DLWP ensemble relative to the other forecasts is better than those for T_2. The DLWP ensemble is again superior to the ECMWF control at both lead times, and more impressively, at weeks 5–6 it is in a statistical tie with the full ECMWF S2S ensemble in the sense that the 95% confidence intervals for the forecasts overlap (black bars in Figure 7d).

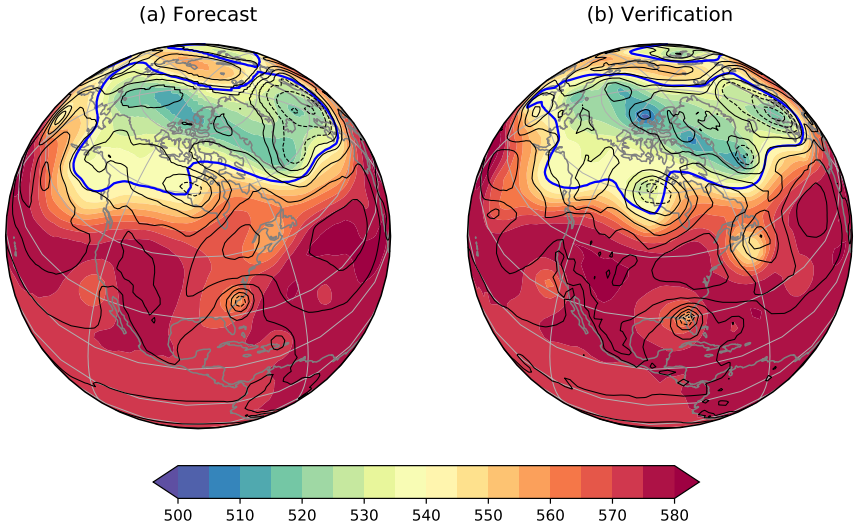

Figure 5. Fields of Z_{500} (color contours, dm) and Z_{1000} (black contours at 100 m intervals with negative values dashed) for (a) a 4.5-day forecast and (b) the verification on 12:00 UTC, September 11, 2017. The blue curve is the 540-dm contour for Z_{500}.

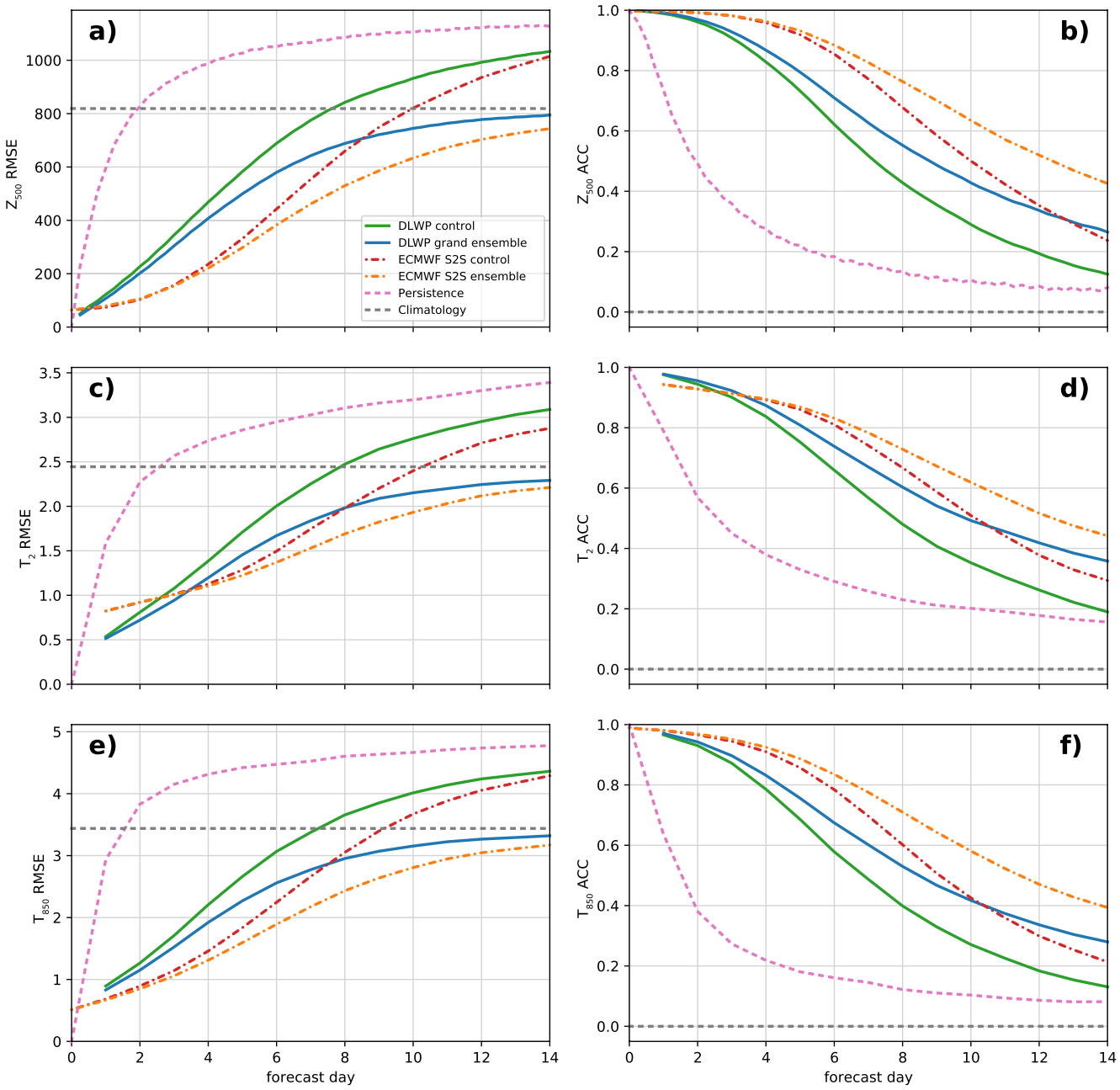

Figure 6. Forecast error as a function of time for the DLWP control member (green) and grand ensemble mean (blue), the ECMWF S2S control (dot-dashed red) and ensemble mean (dot-dashed orange), along with persistence (pink) and climatology (gray) benchmarks for twice-weekly forecasts during 2017–2018. Panels are Z_{500}: (a) RMSE ($m^2\ s^{-2}$) and (b) ACC; daily averaged T_2: (c) RMSE (K) and (d) ACC; T_{850} at 00 UTC: (e) RMSE (K) and (f) ACC. The error is area-weighted in latitude and globally averaged.

ACC scores for the best and worst individual forecasts for each period are shown by the black dots in Figure 7. The ECMWF ensemble has the "best" worst forecasts, or the least susceptibility to bust forecasts, except for 5–6-week T_{850}, for which its worst ACC of −0.22 is worse than the −0.18 value for DLWP ensemble. At weeks 5–6, the best individual forecasts for both the DLWP and ECMWF ensembles have ACC scores of about 0.8 for T_2 and about 0.6 for T_{850}. The globally averaged ACC for the 3–4 week T_2 forecasts exceeds 0.5 roughly 25% of the time for the DLWP grand ensemble and 50% of the time for the ECMWF ensemble.

An example of a good 3–4 week T_2 anomaly forecast for both ensembles is shown in Figure 8. These forecasts were initialized on September 27, 2018. The intensity and spatial variability in the DLWP control

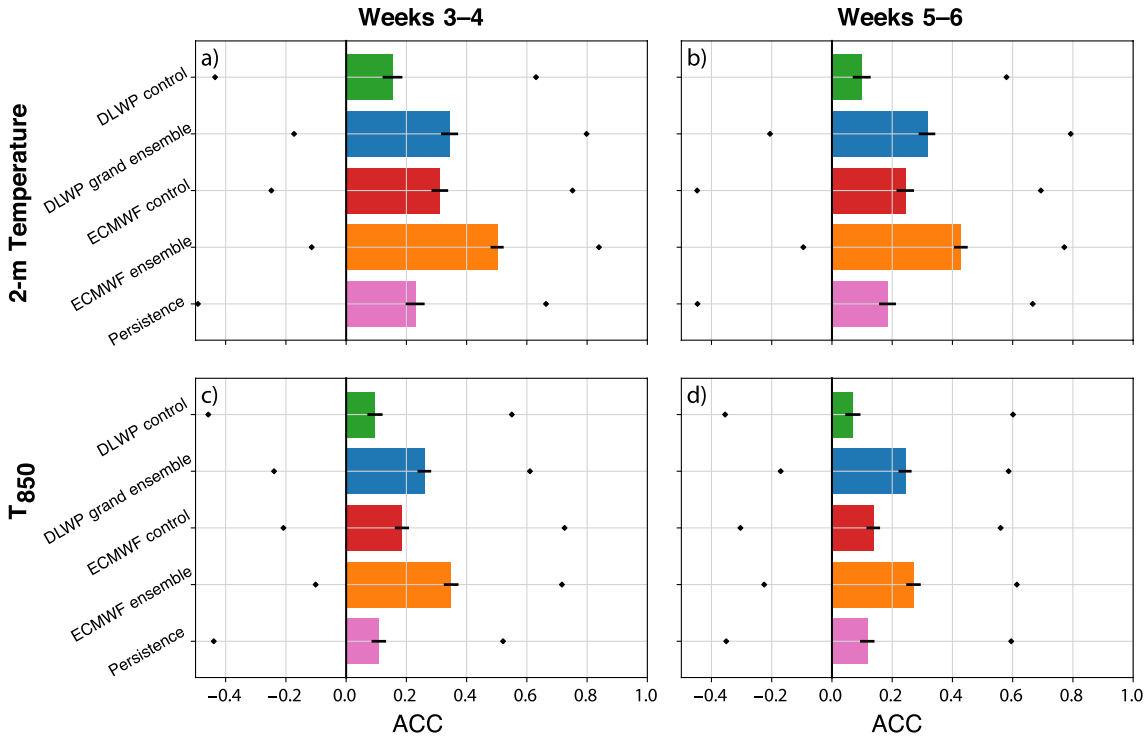

Figure 7. Anomaly correlation coefficient of 2-week-averaged forecasts from the DLWP control (green) and grand ensemble mean (blue), the ECMWF S2S control (red) and ensemble mean (orange), and persistence (pink) for forecasts made twice weekly in 2017–2018. Panels show T_2 for (a) weeks 3–4, (b) weeks 5–6, and T_{850} for (c) weeks 3–4, (d) weeks 5–6. Scores are area-weighted in latitude and globally averaged. Black lines on each bar represent the 95% confidence interval computed using bootstrapping with 10,000 iterations. The black dots show the lowest and highest scores among the 208 forecasts.

forecast are grossly similar to, although modestly weaker and smoother than, those in the ECMWF control, demonstrating that the DLWP forecast is not simply approaching a smooth climatology at long forecast lead times. Similarly, the forecast from the 320-member DLWP grand ensemble exhibits much of the intensity and spatial variability in the 50-member ECMWF ensemble forecast. (Although the horizontal resolution of the ECMWF S2S ensemble members is 31 km beyond lead times of 15 days, the data are archived at coarser resolution and all forecasts are displayed on a 1.5° × 1.5° latitude-longitude grid.)

Anomalies that both verify (Figure 8f) and are common to the DLWP (Figure 8a) and ECMWF S2S (Figure 8b) ensembles include cold in Greenland and warmth in eastern Australia, eastern Siberia and over the adjacent Arctic Ocean. One place where the ECMWF S2S ensemble clearly out performs the DLWP ensemble is over North America, where it better captures the observed large and intense cold anomaly. Another important superiority of the ECMWF ensemble lies in its forecast of a developing El Niño over the equatorial Pacific, although the absence of the El Niño in the DLWP ensemble forecast is not particularly surprising because it does not include any oceanic data.

5.2. Ensemble Probabilistic Scores

Having just examined the ACC scores of the ensemble mean, we now investigate the performance of ensemble-produced probabilistic forecasts, specifically using the continuous ranked probability score (CRPS) and the ranked probability skill score (RPSS).

5.2.1. Continuous Ranked Probability Score

Denoting the ensemble probability distribution function of a forecast for some variable x as $\rho(x)$, the cumulative distribution function (CDF) associated with ρ is

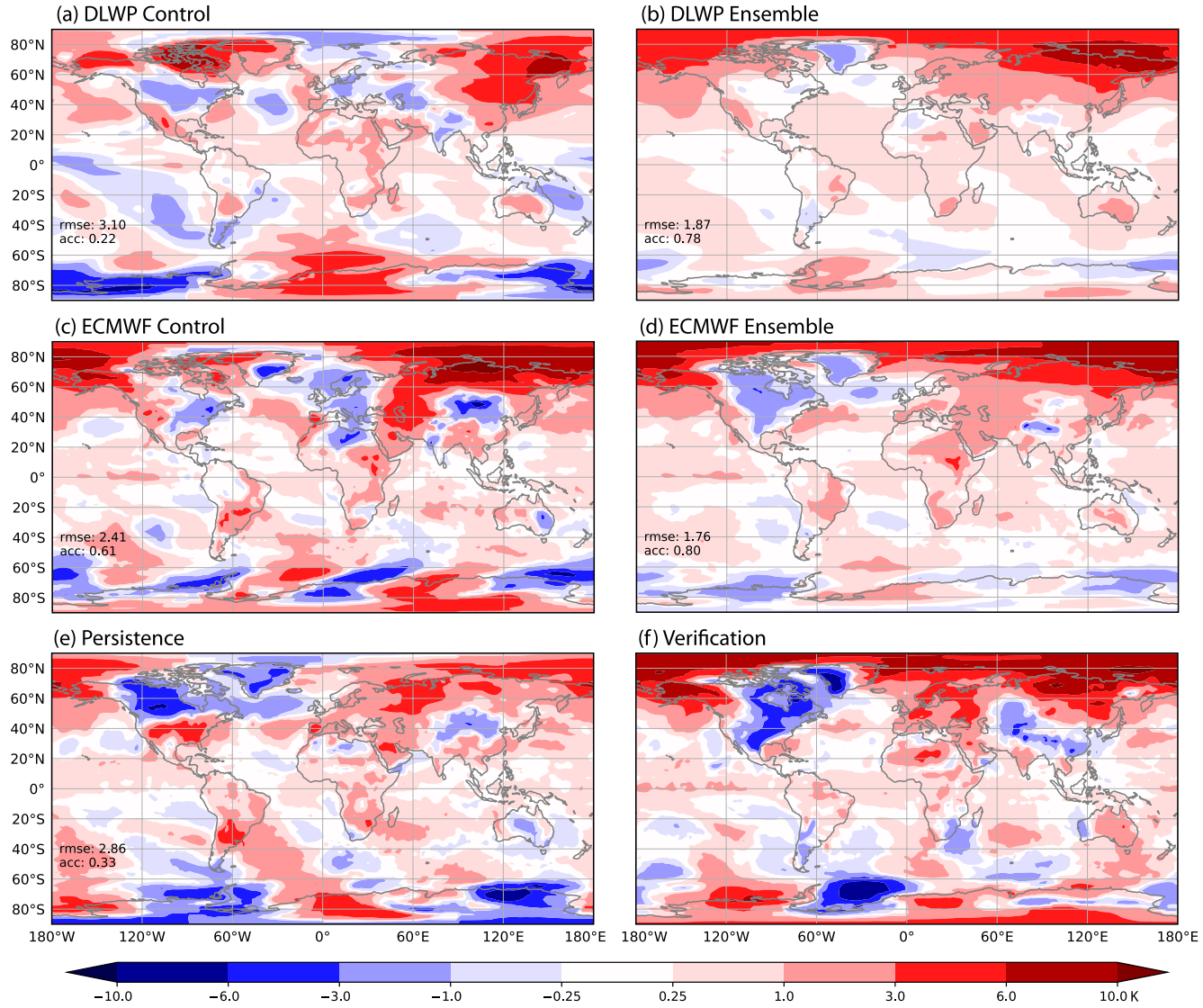

Figure 8. Predicted 2-week average anomalies in T_2 relative to climatology for the period 3–4 weeks after September 27, 2018. Forecasts are from DLWP (a) control and (b) grand ensemble, ECMWF S2S (c) control and (d) ensemble, and (e) persistence; (f) is the verification. The "rmse" and "acc" numbers are global averages; the rmse is the root-mean-squared error of the *anomalies*.

$$P(x) = \int_{-\infty}^{x} \rho(y)dy. \tag{1}$$

If the verifying value occurs at x_a, a CDF for the observation may be defined as

$$P_a(x) = H(x - x_a), \tag{2}$$

where

$$H(s) = \begin{cases} 0 & s < 0 \\ 1 & s \geq 0 \end{cases} \tag{3}$$

is the Heaviside function. The CRPS may then be defined as (Hersbach, 2000):

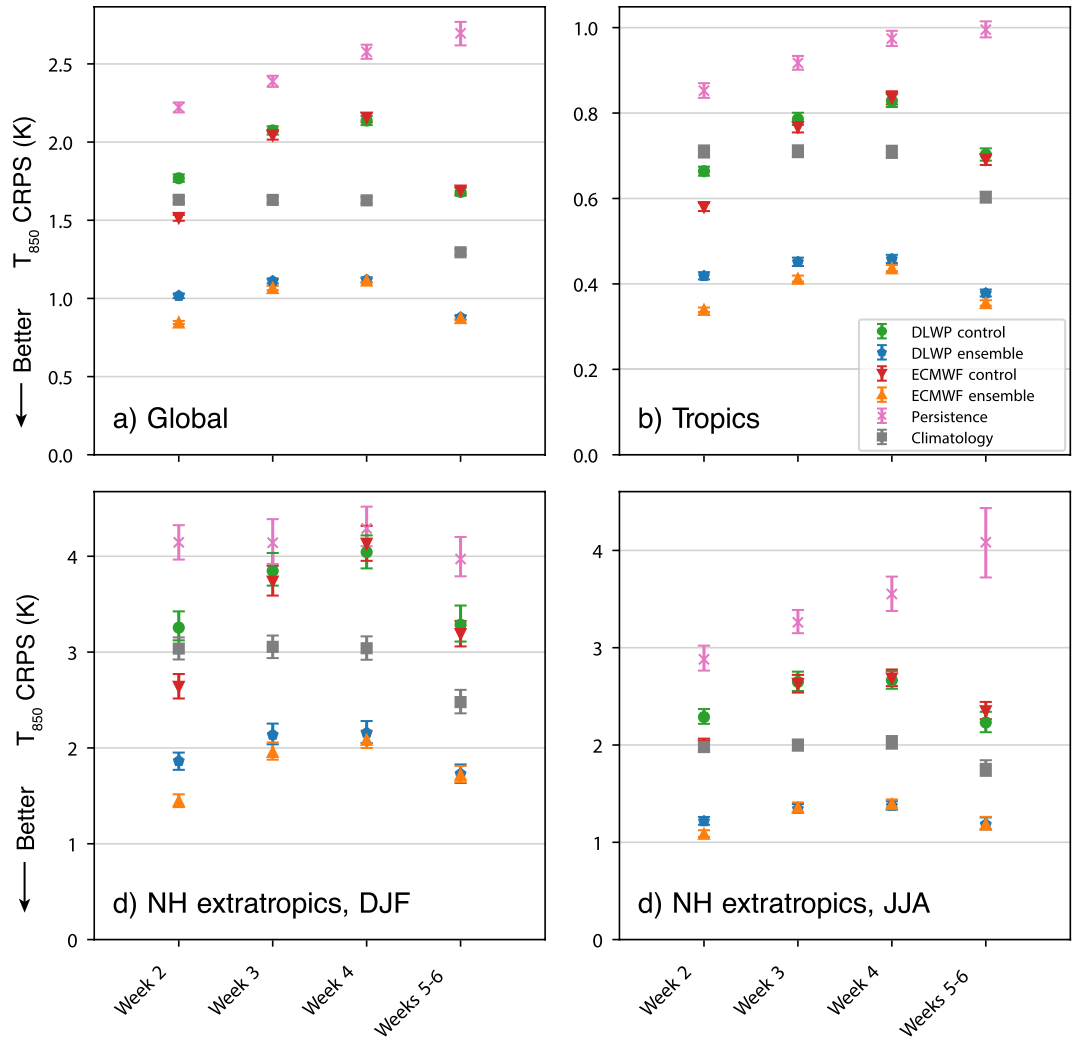

Figure 9. Continuous ranked probability score (CRPS; lower is better) in T_{850} from the DLWP control (green circles), DLWP grand ensemble (blue pentagons), ECMWF S2S control (red downward pointing triangles), ECMWF ensemble (orange upward pointing triangles), persistence (pink crosses), and climatology (gray squares), as a function of averaged forecast lead time. Panels show: (a) Global average, annual mean; (b) average over the tropics (20°S – 20°N), annual mean; (c) average over the northern hemisphere extra-tropics (30°N – 90°N), mean of forecasts initialized in DJF; (d) average over the NH extra-tropics, JJA. Error bars correspond to the 95% confidence interval determined by bootstrapping with 10,000 samples. Note the variations in the scale of the vertical axes.

$$\mathrm{CRPS} = \mathrm{CRPS}(P, x_a) = \int_{-\infty}^{\infty} \left[P(x) - P_a(x)\right]^2 dx. \tag{4}$$

The CRPS score penalizes both overly narrow (confident) forecast distributions that verify incorrectly and overly broad (uncertain) forecast distributions, regardless of the accuracy of the ensemble mean. The CRPS has several desirable properties. First, it is a proper statistical score (Gneiting & Raftery, 2007), meaning that the CRPS is optimized for a forecast which predicts the correct probability distribution of a predicted variable. Second, in contrast to categorical scores, such as the RPSS (see Section 5.2.2), it accounts for information across all possible values of x. Finally, the CRPS reduces to the mean absolute error for a single deterministic forecast, allowing the performance of ensemble and deterministic forecasts to be easily compared.

Figure 9 compares CRPS in T_{850} for forecasts from the DLWP and ECMWF S2S ensembles and control forecasts, along with persistence and climatology benchmarks, averaged over one week for lead times of 2, 3

and 4 weeks, along with a two-week average for a lead time of 5–6 weeks. As with our earlier results for the T_{850} RMSE and ACC of the ensemble mean (Figure 6), the ECMWF ensemble clearly gives better week-2 CRPS scores than the DLWP ensemble. At week 3, when deterministic forecast skill from initial conditions has largely eroded, the ECMWF ensemble continues to give a slightly better global mean result. But by week 4 and weeks 5–6, the DLWP ensemble has caught up and is essentially tied with ECMWF (Figure 9a). At lead times of three weeks or longer, the next best global-mean forecasts are given by climatology, which out-performs the ECMWF and DLWP control forecasts, which are in turn better than persistence. Note that all CRPS scores except persistence improve significantly from week 4 to weeks 5–6 due to the longer two-week averaging window over which slowly evolving, and nominally more predictable, modes make a larger relative contribution.

Focusing on the tropics, from 20°S to 20°N, the CRPS for all models are substantially improved, with numerical values roughly 0.4 times the corresponding global results (Figure 9b). The ECMWF and DLWP ensembles still perform the best, and the relative performance of the various forecast systems is similar to the globally averaged results, although the ECMWF ensemble scores are slightly improved relative to the DLWP ensemble. CRPS scores are worse in the northern hemisphere extra-tropics, 30°N – 90°N (Figures 9c and 9d), and there is a pronounced seasonality in performance. The scores are worse in boreal winter, and the performance of the DLWP ensemble relative to the ECMWF ensemble is also worse. On the other hand, in boreal summer, the CRPS values improve and the DWLP ensemble ties ECMWF in weeks 3, 4 and 5–6. This seasonal difference in extra-tropical performance suggests that the DLWP model performs worse relative to ECMWF when synoptic-scale dynamics exert more influence on the weather.

5.2.2. Ranked Probability Skill Score

The other metric we use to evaluate the ensemble forecasts is the ranked probability skill score (RPSS). To compute the RPSS, K categorical forecasts are first defined. Then let y_i be the probabilistic forecast of the event occurring in category i; let c_i be the climatological probability of the event falling in category i, and bin the verification such that $o_i = 1$ if the event was observed to be in category i and $o_i = 0$ otherwise. The kth components of the cumulative forecast, climatological, and observational distributions Y_k, C_k, and O_k are evaluated for each of the K categories as $Y_k = \sum_{i=1}^{k} y_i$, $C_k = \sum_{i=1}^{k} c_i$, and $O_k = \sum_{i=1}^{k} o_i$.

Ranked probability scores for the forecast (RPS) and climatology (RPS_C) are computed as

$$\mathrm{RPS} = \sum_{k=1}^{K} \left(Y_k - O_k\right)^2 \tag{5}$$

$$\mathrm{RPS}_C = \sum_{k=1}^{K} \left(C_k - O_k\right)^2, \tag{6}$$

and finally, using angled brackets to denote the average over all forecast-observation pairs, the RPSS is defined as

$$\mathrm{RPSS} = 1 - \frac{\langle \mathrm{RPS} \rangle}{\langle \mathrm{RPS}_C \rangle}. \tag{7}$$

The RPS is zero for a perfect forecast and increases positively otherwise, therefore the RPSS for a perfect forecast is one and decreases otherwise. Normalizing $\langle \mathrm{RPS} \rangle$ by $\langle \mathrm{RPS}_C \rangle$ in Equation 7 sets the threshold value below which there is no skill relative to climatology to zero. Like the CRPS, the RPSS is a proper score, but it does depend on the definition of categories. The RPSS is sensitive to the size of the ensemble, having a negative bias for small ensemble sizes (Weigel et al., 2007). Although both the grand ensemble size of 320 and the ensemble size of 50 for ECMWF are large enough to mitigate such bias, in lieu of Equation 7 we will use the de-biased formulation of the RPSS (Weigel et al., 2007), which for ensemble size M is

$$\mathrm{RPSS} = 1 - \frac{\langle \mathrm{RPS} \rangle}{\langle \mathrm{RPS}_C \rangle + D_0 \,/\, M} \tag{8}$$

$$D_0 = \frac{K^2 - 1}{6K}. \quad (9)$$

We compute the RPSS for T_2 and T_{850}, binning into three climatologically equally likely terciles of below-, near-, and above-normal relative to the baseline period of 1981–2010. Tercile bounds for T_2 were determined from daily averaged values (using the times 0, 6, 12, and 18 UTC) for each date in the 30-year record. These daily tercile bounds are then averaged over each one- or two-week verification period. For T_{850}, only the instantaneous 0 UTC values were available from the ECMWF S2S database, therefore a separate climatology is computed from only 0 UTC values to evaluate the ECMWF model (the DLWP forecasts are daily averaged and evaluated using terciles computed from the daily averages). Because each of the forecast categories are, by design, equally likely, the climatological forecast is simply a 33% likelihood of occurrence in each of the categories. Note that because Equations 5 and 6 use cumulative distributions, events verifying in the near-normal category have lower expected random-chance forecast error than events verifying in either the below- or above-normal categories.

Spatially—and temporally-averaged RPSS scores in T_{850} from the DLWP and ECMWF S2S ensembles are shown in Figure 10. The globally-averaged RPSS (Figure 10a) for the DLWP ensemble is well above the zero threshold for random chance at all lead times. Comparing Figures 9 and 10, and recalling that, in contrast to the RPSS, lower CRPS scores are better, both metrics show similar variations in ensemble skill with forecast lead time in all regions and over all time windows. The superiority of the ECMWF ensemble is, nevertheless, greater in the RPSS metric, with a statistically-significant lead over our DLWP ensemble at all forecast lead times and locations, except during JJA in the northern hemisphere extratropics (Figure 10d). The performance difference between the two ensembles is significantly reduced if we consider only locations over land as shown in Figures 10e and 10f, where the skill of the ECMWF ensemble drops significantly after week 2 while the RPSS for the DLWP ensemble, which has no information about SST, remains similar to that over both land and water shown in Figures 10a and 10b. The ability of the model to correctly forecast surface temperatures using the same set of 3×3 filters over both land and ocean highlights the capability of the deep learning approach to capture processes that require complex physical parameterizations in conventional NWP models with no more information than the land-sea mask and the terrain elevation.

The global pattern of RPSS scores, averaged over all of the forecasts in the 2017–2018 test set, is shown by the maps of RPSS scores for both ensembles at weeks 3, 4, and 5–6 in Figure 11. As expected from the plot of global average scores in Figure 10a, the ECMWF ensemble is superior to the DLWP ensemble, with the two becoming more alike and showing more skill in the forecasts averaged over the longer two-week period, weeks 5–6. Particularly at weeks 5–6 (Figures 11e and 11f), the distribution of the low- and high-skill regions in the DLWP and ECMWF ensembles are quite similar. Both ensembles perform almost the same over land and both do very well over the Southern Ocean. The DLWP ensemble shows skill in the tropical oceans, but the ECMWF ensemble does much better in that region, and it largely avoids the loss of skill suffered by the DLWP ensemble over the adjacent subtropical waters. As mentioned previously, our DLWP model likely suffers from the absence of information about SSTs.

The seasonal variation in RPSS at 5–6 weeks is shown in Figure 12. The performance of the DLWP and ECMWF ensembles is most similar in MAM, with global RPSS averages of 0.180 and 0.191, respectively, and analogous spatial patterns of high and low skill. As in the annual mean (Figures 11e and 11f), the skill is relatively high over the tropical and southern oceans. The seasonal averages show more pronounced localized regions of high and no skill (RPSS < 0). One local area where both ensembles show skill, with the DLWP performing best, is in the storm track off the east coast of North America in both DJF and SON. The worst globally averaged RPSS for both ensembles, and the worst performance of DLWP relative to ECMWF, occurs during DJF, but even in this season the spatial patterns of high and low skill are similar.

Finally, we consider surface temperature anomalies, which, as might be expected given the pronounced model drift evident in Figure 4, are significantly improved by bias correction. Maps of RPSS for both the DLWP and ECMWF ensemble forecasts for weeks 5–6 are shown in Figure 13. Bias correction significantly improves the RPSS over land for both ensembles, with much of that improvement in the ECMWF model coming from regions with topography where removing the bias helps correct for differences in the way the

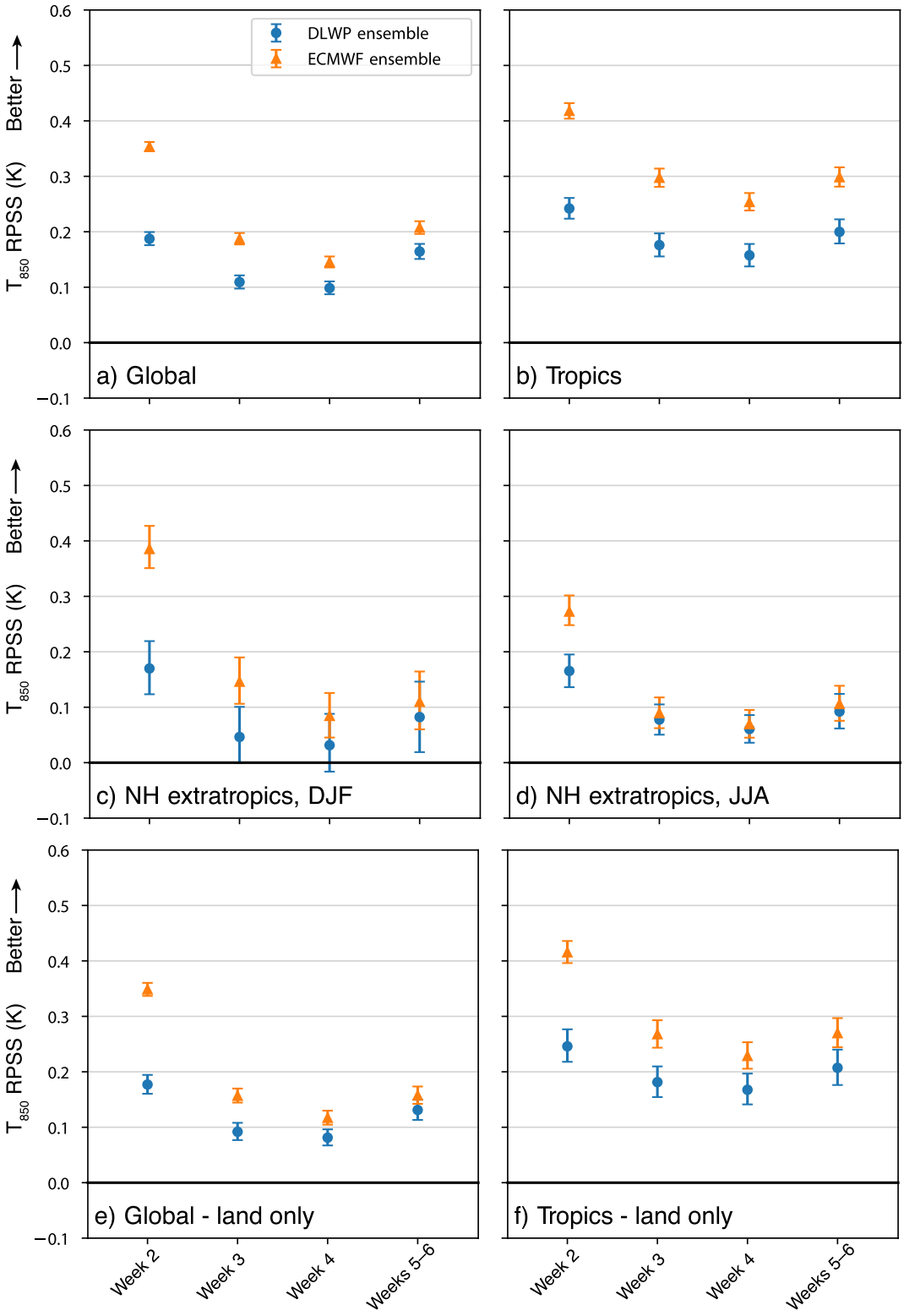

Figure 10. One- or two-week averaged ranked probability skill score (RPSS; higher is better) for T_{850} at indicated forecast lead times. DLWP grand ensemble (blue circles) and the ECMWF S2S ensemble (orange triangles) averaged over the (a) globe, annual mean; (b) tropics (20°S – 20°N), annual mean; (c) northern hemisphere extra-tropics (30°N – 90°N), mean of forecasts initialized in DJF; (d) NH extra-tropics, mean over JJA; (e) and (f), as in (a) and (b) except spatially averaged only over land. Error bars correspond to the 95% confidence interval determined by bootstrapping with 10,000 samples.

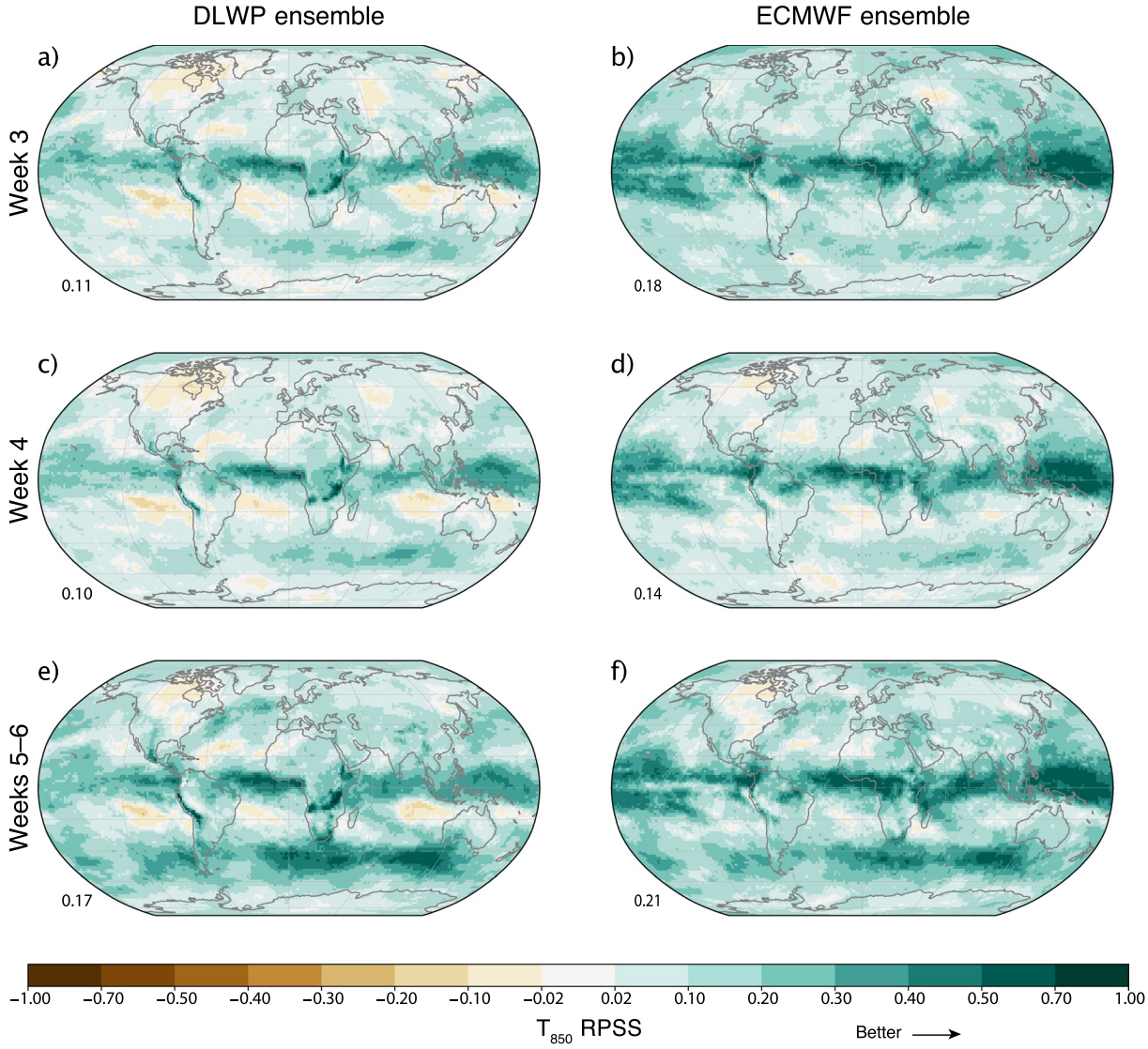

Figure 11. Annual average over the 2017–2018 testing period of T_{850} RPSS scores. Left (right) column show DLWP (ECMWF) ensembles at forecast lead times of (a), (b): 3 weeks, (c), (d) 4 weeks, and (e), (f) 5–6 weeks. The weighted global mean is noted at the lower left in each panel.

orography is represented at different grid resolutions. The topography does not introduce significant bias above the surface; for example the bias-uncorrected forecasts of T_{850} show no trace of a topographic signature and almost no regions in which the RPSS is less than -0.10 (not shown). Bias removal also makes large improvements in T_2 over the oceans in the DLWP ensemble, perhaps partly compensating for the lack of SST data. The regions of highest skill in the bias-corrected forecasts from both ensembles are mostly over the oceans, particularly in the tropics. The global mean RPSS for both ensembles, 0.155 for DLWP and 0.287 for ECMWF, are non-negative, indicating modest skill relative to climatology. The results for the bias-corrected ECMWF ensemble (Figure 4d) are generally consistent with the distribution of RPSS scores from an earlier version of the ECMWF ensemble in (Weigel et al., 2008), which showed skill predominantly over tropical oceans after week 2, despite accounting for bias correction in the forecasts. Improvements in the ECMWF model since (Weigel et al., 2008) have, nevertheless, led to generally higher RPSS values over the oceans and much better performance over the Western Pacific and Indian Oceans, as would be apparent in a comparison of these results with their Figure 4, although such comparison must be qualified by noting the current forecasts and those in (Weigel et al., 2008) verify in different years.

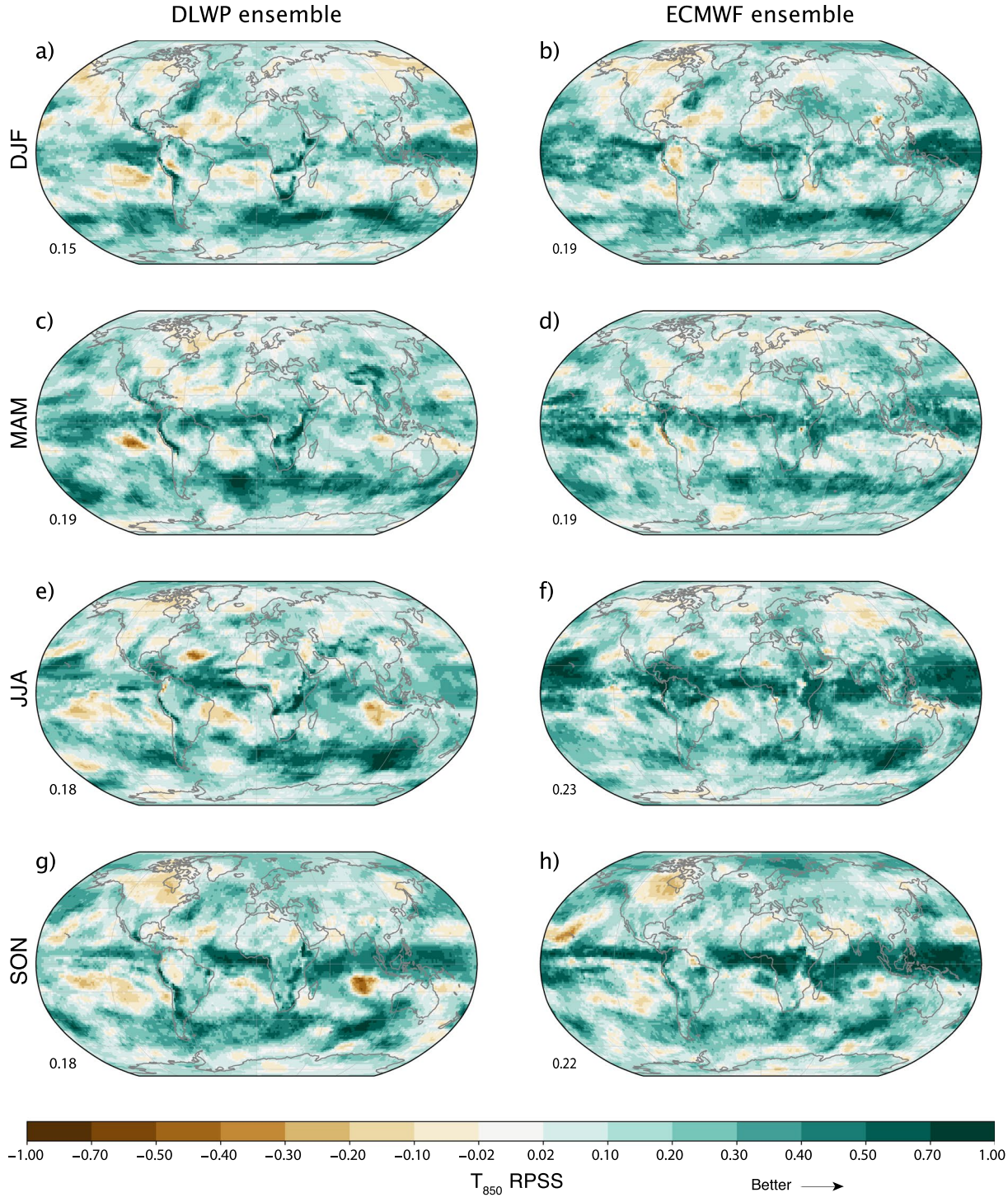

Figure 12. Maps of seasonally-averaged RPSS scores in T_{850} during the testing period of 2017–2018. Left (right) column DLWP (ECMWF) ensembles at 5–6 weeks forecast lead time for months (a), (b): DJF, (c), (d) MAM, (e), (f) JJA, and (g), (h) SON. The weighted global mean is noted at the lower left in each panel.

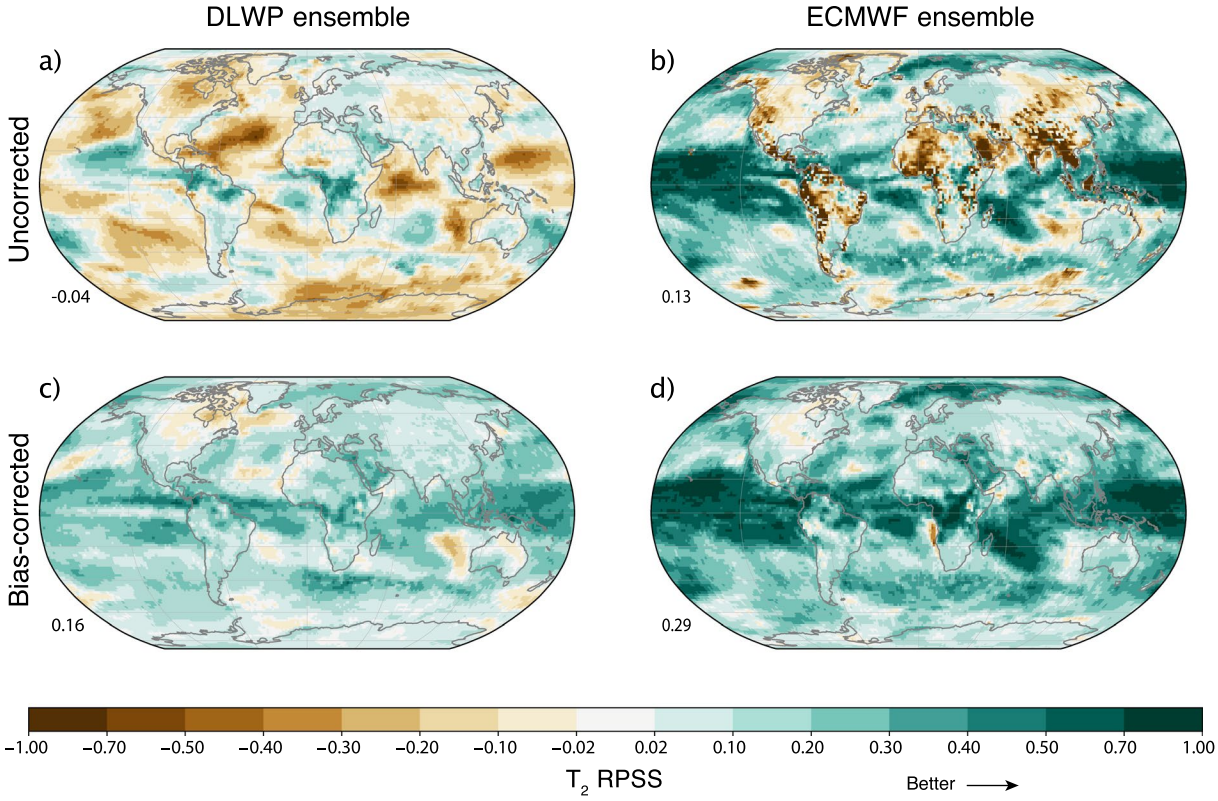

Figure 13. Annual average RPSS skill maps for T_2 at weeks 5–6. Without bias correction: (a) DLWP ensemble, (b) ECMWF ensemble; with bias correction: (c) DLWP ensemble, (d) ECMWF ensemble. The weighted global mean is noted at the lower left in each panel.

The tendency of the 5–6 week bias-corrected T_2 ECMWF ensemble forecast to perform similarly to the DLWP forecast over land, and better over the oceans, is again apparent in the seasonal results for SON shown in Figure 14. Note in particular, the negative RPSS score over the equatorial eastern Pacific Ocean, which arises because the DLWP ensemble fails to correctly capture the onset of a weak El Niño event in 2018. In contrast, the ECMWF ensemble, with coupling to an ocean model, exhibits high skill throughout the same region.

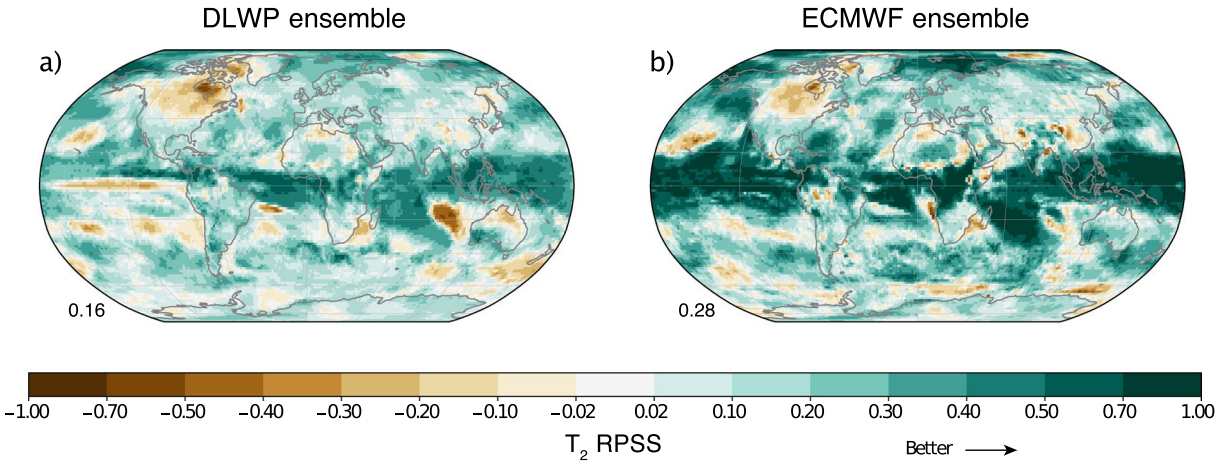

Figure 14. As in Figure 13, but only for bias-corrected forecasts initialized in SON.

6. Conclusions

As a first step toward developing a deep-learning-based ensemble system for S2S forecasting, we refined our previous data-driven global model (Weyn et al., 2020) by improving the resolution at the equator to approximately 1.4°by 1.4°and by adding two more physical fields, the temperature at 850 hPa and total column water vapor. These refinements allowed the model to both spontaneously generate tropical cyclones and also produce a reasonable, though far from state-of-the-art, 4.5-day deterministic forecast of Hurricane Irma. Despite the higher resolution and the expansion from four to six spherical shells of prognostic variables, the model remains very computationally efficient. A one-week forecast, stepped forward with a 12-h time step (and 6-h resolution), can be performed in approximately 1/10th of a second on an Nvidia Tesla V100 graphics processing unit (GPU).

We exploited this efficiency to generate large ensemble forecasts. Only about 3 min are required to produce a 320-member six-week ensemble forecast. Those 320 ensemble members were generated by running 32 different DLWP models, trained to slightly different convolutional filter coefficients, on each of 10 initial conditions. The initial conditions were non-optimal; rather than including information from singular vectors, they were simply drawn from the ERA5 archive. The strategy of training DLWP models with slightly different filter coefficients was, on the other hand, very effective, adding significant skill to the ensemble mean and greatly increasing the ensemble spread (Figures 2 and 6). Our DLWP model requires 6–8 days of computation to train on a single Tesla V100 GPU. We were able to economize by training only eight of our 32 multi-model ensemble members from scratch with different initial seeds and filling out the ensemble with models using filter coefficients from different checkpoints saved during the training cycles.

As was also the case for the ECMWF S2S forecast, the DLWP ensemble mean forecast was a significant improvement over that from a single control member. In particular, the average RMSE over the 2017–2018 test set of DLWP ensemble forecasts for Z_{500}, T_{850}, and 2-m temperature remained below climatology for at least 14 days, while the anomaly correlation coefficients remained above 0.6 for 7–8 days. Not surprisingly the ECMWF S2S ensemble did perform better, particularly at earlier lead times, and it also gave ACC scores exceeding the 0.6 threshold out to 10 days. At longer lead times, week 3–4 or week 5–6 averages, the ACC scores of the ECMWF and DLWP ensemble means were positive and better than persistence, but still relatively low. The 2017–2018 averaged scores ranged from roughly 0.25 to 0.5, with the ECMWF ensemble performing better in all cases except for T_{850} at weeks 5–6, for which both ensembles were in a statistical tie with an ACC of approximately 0.25.

We examined two probabilistic measures of ensemble skill, the continuous ranked probability score (CRPS) and the ranked probability skill score (RPSS). For the CRPS, the DLWP and ECMWF S2S ensembles produce essentially the same week-4 and weeks-5–6 scores. At shorter lead times the ECMWF ensemble is superior, performing marginally better at week 3 and distinctly better than the DLWP ensemble at week 2. Both the DLWP and ECMWF ensembles clearly out-performed climatology and persistence. Examining seasonal and regional contrasts showed that in the northern hemisphere extra-tropics, the DLWP ensemble performed best, and on par with the ECMWF ensemble in the summer season, while performing worst in winter.

Like the CRPS, the spatially and temporally averaged RPSS scores showed modest skill relative to climatology at all lead times. The ECMWF ensemble RPSS scores exceeded those of the DWLP ensemble by larger margins than in the CRPS metric, except in summer in the northern extra-tropics when both ensembles again achieved similar scores. In both the globally-averaged and tropics-only-averaged RPSS, the differential by which the ECMWF RPSS score exceeds that of DLWP is smaller over land than over the full globe. Global maps comparing the ECMWF and DLWP RPSS scores show generally similar regions of higher and lower skill, except that the ECMWF ensemble performs better over the tropical oceans. At weeks 5–6, the spatial distribution of regions of skill and no-skill in the RPSS metric over land are surprisingly similar between the ECMWF and DLWP ensembles (Figure 12). One reason the DLWP model performs poorer over the tropical oceans is likely due to its lack of SST data, as suggested by its failure in the eastern equatorial Pacific during the onset of a weak El Niño event in 2018.

Although our current data-driven DLWP model is worse than operational NWP models for the deterministic prediction of synoptic-scale weather patterns, its capability to learn physics-based phenomena, including

the complex evolution of near-surface temperatures and long-term patterns in the convection-dominated tropics, is remarkable. As such, DLWP may prove a valuable tool for supplementing NWP-based S2S forecasts where they are weakest: in the tropics and in the spring and summer months.

There are many avenues for further development of our elementary DLWP ensemble system. One obvious shortcoming is that our DLWP model does not yet forecast precipitation. This might be addressed by adding precipitation to the current set of six prognostic 2D fields that are recursively stepped forward by the model, although instead of including the precipitation at previous times in the CNN architecture, it could alternatively be diagnosed from the other fields after each step (Larraondo et al., 2019).

The DLWP model's computational efficiency can be used for more than simply producing timely operational forecasts; it also enables researchers to make unprecedented use of large numbers of reforecasts for past weather events. For our bias correction, we computed 85,800 reforecasts in a matter of hours on a single GPU. We only used these reforecasts to correct the average drift in the DLWP model, but one could also use them to calibrate ensemble probability distributions, analyze model errors, or investigate the sources of predictability captured by the model. Historically, adjoint models (e.g., Doyle et al., 2014), which are tangent linear, differentiable approximations to full non-linear dynamical NWP models, have been used to examine how model errors depend on initial condition uncertainties (for example, whether errors in moisture over the US have a strong influence on the location or intensity of a cyclone over Europe). Adjoint models are difficult to create for complex operational NWP models. Yet, because a CNN is fully differentiable, it is easy to produce the corresponding adjoint model, enabling studies of error growth and atmospheric predictability. Recent work on the interpretation of deep neural networks may provide some valuable tools for this form of analysis (Ebert-Uphoff & Hilburn, 2020; Toms et al., 2020).

Data Availability Statement

The ERA5 reanalysis data are available via the Copernicus Climate Data Store (http://doi.org/10.24381/cds.bd0915c6 and http://doi.org/10.24381/cds.adbb2d47). The ECMWF S2S forecasts are available at https://apps.ecmwf.int/datasets/data/s2s. The T42 and T63 IFS forecasts are available from Rasp et al. (2020b).

Acknowledgments
The authors thank Elizabeth Barnes and Peter Dueben for comments that helped improve the paper. Jonathan A. Weyn and Dale R. Durran's contributions to this research were funded by Grant N00014-17-1-2660 from the Office of Naval Research (ONR). Dale R. Durran was also supported by Grant N00014-20-1-2387 from ONR. Jonathan A. Weyn was also supported by a National Defense Science and Engineering Graduate (NDSEG) fellowship from the Department of Defense (DoD). Computational resources were provided by Microsoft Azure via a grant from Microsoft's AI for Earth program. Stan Posey (Nvidia) also generously donated two V100 GPUs to the University of Washington which were used for this work.

References

Adames, Á. F., & Kim, D. (2016). The MJO as a dispersive, convectively coupled moisture wave: Theory and observations. *Journal of the Atmospheric Sciences*, *73*(3), 913–941. https://doi.org/10.1175/JAS-D-15-0170.1

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Buizza, R. (2019). Introduction to the special issue on "25 years of ensemble forecasting". *Quarterly Journal of the Royal Meteorological Society*, *145*(S1), 1–11. https://doi.org/10.1002/qj.3370

Doyle, J. D., Amerault, C., Reynolds, C. A., & Reinecke, P. A. (2014). Initial condition sensitivity and predictability of a severe extratropical cyclone using a moist adjoint. *Monthly Weather Review*, *142*(1), 320–342. https://doi.org/10.1175/MWR-D-13-00201.1

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

Ebert-Uphoff, I., & Hilburn, K. A. (2020). *Evaluation, tuning and interpretation of neural networks for meteorological applications*. http://arxiv.org/abs/2005.03126

Fortin, V., Abaza, M., Anctil, F., & Turcotte, R. (2014). Why should ensemble spread match the rmse of the ensemble mean? *Journal of Hydrometeorology*, *15*(4), 1708–1713. https://doi.org/10.1175/JHM-D-14-0008.1

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)0150559:DOTCRP2.0.CO;2

Hwang, J., Orenstein, P., Cohen, J., Pfeiffer, K., & Mackey, L. (2019). Improving subseasonal forecasting in the western U.S. With machine learning. *In Proceedings of the acm sigkdd international conference on knowledge discovery and data mining*. (pp. 2325–2335). https://doi.org/10.1145/3292500.3330674

Isaksen, L., Bonavita, M., Buizza, R., Fisher, M., Haseler, J., Leutbecher, M., & Raynaud, L. (2010). Ensemble of data assimilations at ECMWF. *ECMWF Technical Memoranda*, *636*, 1–41.

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. http://arxiv.org/abs/1412.6980

Larraondo, P. R., Renzullo, L. J., Inza, I., & Lozano, J. A. (2019). *A data-driven approach to precipitation parameterizations using convolutional encoder-decoder neural networks*. http://arxiv.org/abs/1903.10274

Leutbecher, M. (2018). Ensemble size: How suboptimal is less than infinity? *Quarterly Journal of the Royal Meteorological Society*, *145*(S1), 107–128. https://doi.org/10.1002/qj.3387

Mayer, K. J., & Barnes, E. A. (2020). Subseasonal forecasts of opportunity identified by an interpretable neural network. *Earth and Space Science Open Archive*. https://doi.org/10.1002/essoar.10505448.1

Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C., & Liniger, M. A. (2018). Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *Journal of Geophysical Research: Atmospheres*. https://doi.org/10.1029/2017JD027923

Palmer, T. (2018). The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*. https://doi.org/10.1002/qj.3383

Rasp, S., & 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*, e2020MS002405. https://doi.org/10.1029/2020MS002405

Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *126*(563), 649–667. https://doi.org/10.1002/qj.49712656313

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *In Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. (Vol. 9351, pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Scher, S., & Messori, G. (2019). Weather and climate forecasting with neural networks: Using GCMs with different complexity as study-ground. *Geoscientific Model Development Discussions*, 1–15. https://doi.org/10.5194/gmd-2019-53

Scher, S., & Messori, G. (2020). *Ensemble neural network forecasts with singular value decomposition*. http://arxiv.org/abs/2002.05398

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002. https://doi.org/10.1029/2019MS002002

Ullrich, P. A., Devendran, D., & Johansen, H. (2016). Arbitrary-order conservative and consistent remapping and a theory of linear maps: Part II. *Monthly Weather Review*, *144*(4), 1529–1549. https://doi.org/10.1175/MWR-D-15-0301.1

Ullrich, P. A., & Taylor, M. A. (2015). Arbitrary-order conservative and consistent remapping and a theory of linear maps: Part I. *Monthly Weather Review*, *143*(6), 2419–2440. https://doi.org/10.1175/MWR-D-14-00343.1

Vitart, F. (2004). Monthly forecasting at ECMWF. *Monthly Weather Review*, *132*(12), 2761–2779. https://doi.org/10.1175/MWR2826.1

Vitart, F. (2014). Evolution of ECMWF sub-seasonal forecast skill scores. *Quarterly Journal of the Royal Meteorological Society*, *140*(683), 1889–1899. https://doi.org/10.1002/qj.2256

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., et al. (2017). The Subseasonal to Seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, *98*(1), 163–173. https://doi.org/10.1175/BAMS-D-16-0017.1

Weigel, A. P., Baggenstos, D., Liniger, M. A., Vitart, F., & Appenzeller, C. (2008). Probabilistic verification of monthly temperature forecasts. *Monthly Weather Review*, *136*(12), 5162–5182. https://doi.org/10.1175/2008mwr2551.1

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2007). The discrete Brier and ranked probability skill scores. *Monthly Weather Review*, *135*(1), 118–124. https://doi.org/10.1175/MWR3280.1

Weyn, J. A., Durran, D. R., & 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. https://doi.org/10.1029/2019ms001705

Weyn, J. A., Durran, D. R., & 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. https://doi.org/10.1029/2020MS002109

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J. T., Lazo, J. K., Kumar, A., et al. (2017). Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological Applications*, *24*(3), 315–325. https://doi.org/10.1002/met.1654