BAMS Article

Skill of Medium-Range Forecast Models Using the Same Initial Conditions

L. Magnusson, D. Ackerley, Y. Bouteloup, J.-H. Chen, J. Doyle, P. Earnshaw, Y. C. Kwon, M. Köhler, S. T. K Lang, Y.-J. Lim, M. Matsueda, T. Matsunobu, R. McTaggart-Cowan, A. Reinecke, M. Yamaguchi, and L. Zhou

ABSTRACT: In the Different Models, Same Initial Conditions (DIMOSIC) project, forecasts from different global medium-range forecast models have been created based on the same initial conditions. The dataset consists of 10-day deterministic forecasts from seven models and includes 122 forecast dates spanning one calendar year. All forecasts are initialized from the same ECMWF operational analyses to minimize the differences due to initialization. The models are run at or near their respective operational resolutions to explore similarities and differences between operational global forecast models. The main aims of this study are 1) to evaluate the forecast skill and how it depends on model formulation, 2) to assess systematic differences and errors at short lead times, 3) to compare multimodel ensemble spread to model uncertainty schemes, and 4) to identify models that generate similar solutions. Our results show that all models in this study are capable of producing high-quality forecasts given a high-quality analysis. But at the same time, we find a large variety in model biases, both in terms of temperature errors and precipitation. We are able to identify models whose forecasts are more similar to each other than they are to those of other systems, due to the use of similar model physics packages. However, in terms of multimodel ensemble spread, our results also demonstrate that forecast sensitivities to different model formulations still are substantial. We therefore believe that the diversity in model design that stems from parallel development efforts at global modeling centers around the world remains valuable for future progress in the numerical weather prediction community.

KEYWORDS: Model comparison; Model errors; Model evaluation/performance; Numerical weather prediction/forecasting

https://doi.org/10.1175/BAMS-D-21-0234.1

Corresponding author: Linus Magnusson, linus.magnusson@ecmwf.int

In final form 15 June 2022

AFFILIATIONS: Magnusson and Lang—European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom; **Ackerley and Earnshaw**—Met Office, Exeter, United Kingdom; **Bouteloup**—Météo-France, Toulouse, France; **Chen**—NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey; **Doyle and Reinecke**—U.S. Naval Research Laboratory, Monterey, California; **Kwon and Lim**—Korea Meteorological Administration, Seoul, South Korea; **Köhler**—Deutscher Wetterdienst, Offenbach, Germany; **Matsueda**—University of Tsukuba, Tsukuba, Japan; **Matsunobu**—University of Tsukuba, Tsukuba, Japan, and Ludwigs-Maximilians-Universität, Munich, Germany; **McTaggart-Cowan**—Environment and Climate Change, Canada, Montreal, Quebec, Canada; **Yamaguchi**—Japan Meteorological Agency, Tokyo, Japan; **Zhou**—Atmospheric and Oceanic Sciences Program, Princeton University, Princeton, New Jersey

Global medium-range numerical weather predictions (NWP) are generated daily by several modeling centers, each of which develops largely independent forecast systems (observation handling, data assimilation, and the forecast model; for acronyms of centers and models, see Table 1). However, a continuous exchange of knowledge via conferences, scientific publications, and staff movements results in NWP systems that share similar fundamental concepts, for example, in dynamical cores and physical parameterization schemes. From this the main question for the DIMOSIC (Different Models, Same Initial Conditions) project arises: are we approaching a point where all models yield forecasts that are statistically similar?

Errors in NWP forecasts originate from uncertainty in the initial conditions and imperfections in the models. To investigate local (in time and space) error growth due to model differences, initial condition differences need to be minimized and short lead times need to be evaluated to minimize accumulated chaotic error growth [see Dalcher and Kalnay (1987) for a conceptual model about error growth]. One possible approach is to initialize different models from the same initial conditions, assuming that the effect of the initial shock from nonnative initial conditions is small. To understand biases in climate models, this approach has been adopted in the past by the Transpose-AMIP project, where climate models were initialized from the same analyses and evaluated on a medium-range forecast time scale (Williams et al. 2013). In a collaboration between GFDL and ECMWF, the global FV3 model was initialized from both NCEP

Table 1. Acronyms for institutes and models.

Acronym	Expansion
Arpege	Action de Recherche Petite Echelle Grande Echelle
CMC	Canadian Meteorological Centre
DWD	Deutscher Wetterdienst
ECMWF	European Centre for Medium-Range Weather Forecasts
FV3	Finite-Volume Cubed-Sphere Dynamical Core
GEM	Global Environmental Multiscale Model
GFDL	Geophysical Fluid Dynamics Laboratory
GFS	Global Forecast System
GSM	Global Spectrum Model
ICON	Icosahedral Nonhydrostatic Model
IFS	Integrated Forecasting System
JMA	Japan Meteorological Agency
KMA	Korea Meteorological Administration
NCEP	National Centers for Environmental Prediction
NRL	Naval Research Laboratory
SHiELD	System for High-Resolution Prediction on Earth-to-Local Domains
UM	Unified Model

and ECMWF initial conditions. Valuable results emerged from this intercomparison, both in terms of midlatitude forecast errors (Magnusson et al. 2019) and tropical cyclone forecast errors (Chen et al. 2019).

The same approach has also been adopted in the Dynamics of the Progress in Earth and Planetary Science Open Access Atmospheric General Circulation Modeled On Nonhydrostatic Domains (DYAMOND) project (Stevens et al. 2019; Judt et al. 2021). In this project, the ECMWF global 9-km meteorological analysis of 1 August 2016 was used to initialize nine global storm-resolving models (GSRMs) with a grid spacing of 5 km or less. Then, deterministic 40-day integrations from each model were analyzed, with a focus on energy budgets, precipitation, and tropical cyclones. This project showed the utility of understanding intermodel differences and assessing the sensitivity of results to a particular implementation. Therefore, following a complementary method to DYAMOND, one can use a set of models at a relatively lower resolution to perform multiple medium-range simulations and analyze the characteristics of these medium-range predictions. This idea promotes the form of the DIMOSIC project.

The main aims of DIMOSIC are the following:

- Identify and understand relationships between model formulation and forecast skill. By starting different models from the same initial conditions and comparing forecast skill for different regions, the strengths and weaknesses of individual models can be identified.
- Assess systematic differences and errors at short lead times. The DIMOSIC protocol eliminates short-range biases that originate from the use of different data assimilation systems. Any rapid development of distinct biases is therefore an indication of systematic differences between the formulations of the forecast models.
- Compare multimodel ensemble spread to model uncertainty schemes. This may highlight areas where current model uncertainty schemes do not fully represent error sources related to model formulation.
- Identify models that generate similar solutions. This information will help forecasters and end users to appreciate the significance of differences between guidance from different models. It will also help developers to identify systematic errors associated with specific model components.

The institutes (models, key reference) represented in this report are DWD (ICON, https://code.mpimet.mpg.de/projects/iconpublic), Météo-France (Arpege; Roehrig et al. 2020), ECMWF (IFS; ECMWF 2020), the Met Office (UM; Walters et al. 2019), Environment and Climate Change Canada (CMC-GEM; https://github.com/ECCC-ASTD-MRD/gem), GFDL (SHiELD; Harris et al. 2020), and JMA (JMA-GSM; JMA 2019).

Because most of the contributing modeling centers have experience with initializing their models with ECMWF operational analyses or reanalyses (ERA-Interim or ERA5), this became the choice for the common initial conditions. For each model, the same set of 10-day forecasts are initialized every third day for a 1-yr period (6 June 2018–4 June 2019). The model grid spacings are the same or similar to what is used operationally at each institute.

We begin this paper by summarizing the contributing models and the compilation of the dataset. This section also describes the ingestion of the ECMWF initial conditions in each model and any adjustments that were required. Results are then presented in the form of forecast verification scores and mean errors (biases), augmented by an assessment of multimodel ensemble spread and similarities between models. The project outcomes and progress made toward achieving DIMOSIC objectives are discussed in the final section.

Models and data

Model descriptions. All models that participate in the DIMOSIC project are global models used for medium-range weather forecasting. Table 2 provides a brief overview of the configurations of each model in terms of grid spacing, dynamical core, and selected physical parameterizations. The table also provides key references for each model that contain additional details about specific components. For the horizontal grid spacing, Arpege has a horizontally varying grid spacing with refinement to 5 km over Europe.

Although all models have grid spacings between 5 and 25 km, they are in different stages of implementing methods believed to be important for kilometer-scale forecasts. The UM, SHiELD, and ICON dynamical cores are nonhydrostatic, while others are hydrostatic. The SHiELD and ICON models have finite-volume dynamical cores. For the horizontal discretization, IFS, Arpege, and JMA-GSM use spectral methods, while the other models use variants of a gridpoint model. Another important part for the dynamical core is the time stepping (Mengaldo et al. 2019), where the models with a finite-volume core use a combination of explicit and semi-implicit time stepping while the other models use semi-implicit or fully implicit time stepping.

For the large-scale precipitation and cloud processes, most models use a single-moment microphysics parameterization, but with a different number of categories for cloud condensate. For the convection parameterization, IFS, ICON, and Arpege (the version used in this study) use the Tiedtke–Bechtold convection scheme (Tiedtke 1993; Bechtold et al. 2008). It is worth noting that even if model components are built on the same original idea, the implementation

Table 2. Model descriptions. FE = Finite element, FD = finite difference, FV = finite volume, H = hydrostatic, NH = nonhydrostatic, CLWC = cloud liquid water content, CIWC = cloud ice water content, CRWC = cloud rainwater content, CSWC = cloud snow water content, TKE = turbulent kinetic energy scheme, ED(MF) = eddy diffusivity (mass flux), RRTM = Rapid Radiative Transfer Model.

Model	Institute	Version	Resolution	Dynamical core	Convection	Cloud water content cat.	Turbulence	Orographic drag	Radiation	Key reference(s)
IFS	ECMWF	47r1	9 km/137 levels	Spectral/FE/H	Tiedtke–Bechtold	CLWC, CIWC, CRWC, CSWC	EDMF	Lott–Miller	RRTM (EcRad)	ECMWF (2020)
CMC-GEM	CMC	v5.0.2	15 km/80 levels	Yin–Yang grid/FD/H	Kain and Fritsch (deep) + Bechtold (shallow)	Single liquid ice	TKE	Lott–Miller	Correlated K	Girard et al. (2014) McTaggart-Cowan et al. (2019)
ARPEGE	Météo-France	46T1	5–25 km/105 levels	Spectral/FE, H	Tiedtke–Bechtold + shallow mass flux	CLWC, CIWC, CRWC, CSWC	TKE	Lott–Miller	RRTM	Roehrig et al. (2020)
UM	Met Office	—	10 km/70 levels	Regular lon–lat grid/FD, NH	Gregory and Rowntree mass flux	Liquid and ice mixing ratio	First-order turbulence closure	Spectral subgrid	SOCRATES	Walters et al. (2019)
SHIELD	GFDL	—	13 km/91 levels	Cube-sphere/FV, NH	Simplified Arakawa–Schubert	CLWC, CIWC, CRWC, CSWC, graupel	Yonsei Uni.	Lott–Miller and Kim–Doyle	RRTM	Harris et al. (2020)
ICON	DWD	21 April 2021	13 km/90 levels	Icosaheder/FV, NH	Tiedtke–Bechtold	CLWC, CIWC, CRWC, CSWC	TKE	Lott–Miller	RRTM (EcRad)	DWD (2022)
JMA-GSM	JMA	GSM1705	20 km/100 levels	Spectral/FD, H	Simplified Arakawa–Schubert	Cloud water content	Hybrid TKE and ED	Type A and B scheme	Two stream approx.	JMA (2019)

can vary substantially. Bengtsson et al. (2019) indicated that the Tiedtke–Bechtold convection scheme needed significant retuning to work properly with the other model components such as the microphysics scheme in the FV3GFS model. The convection and microphysics schemes were revised in IFS model cycle 47r3 and included improvements to the interaction between the schemes (Bechtold et al. 2020). The impact of this upgrade is evaluated in this paper to give an example of an incremental change in one model.

The planetary boundary layer is represented in the Arpege, ICON, and CMC-GEM models by using different formulations of turbulent kinetic energy (TKE) closures, while JMA-GSM uses a hybrid eddy diffusivity–TKE scheme. Several models use the Lott and Miller (1997) orographic drag parameterization, one of several subgrid-scale orography schemes reviewed by van Niekerk et al. (2020). All models except Arpege include a parameterization scheme for nonorographic gravity wave drag.

For the sea surface temperature (SST) evolution, all but three (IFS, SHiELD, and CMC-GEM) models use persistent anomalies from the analysis. For the exceptions, IFS uses a partial coupling to the 3D ocean NEMO model (Mogensen et al. 2017), SHiELD is coupled with a 1D mixed layer ocean model (Pollard et al. 1973), and CMC-GEM uses the initializing analysis and a thermodynamic mixed layer ocean model (Zeng and Beljaars 2005).

Data processing and verification data. The forecast output was interpolated to a regular 0.25° grid at each contributing center followed by an interpolation to a common 0.5° grid using an average interpolation method. This step was necessary to standardize the final interpolation step as other modeling centers may have used different methods to produce the 0.25° grid.

In most of the results presented here, the forecasts are verified against multicenter analyses based on the mean of ECMWF, MetOffice, NCEP, CMC, KMA, and JMA analyses available in the The International Grand Global Ensemble (TIGGE) archive (Swinbank et al. 2016). The multianalysis ensemble spread from TIGGE was evaluated in, e.g., Bauer et al. (2016). Note that some analyses are missing in the TIGGE archive during the verification period from individual centers, and for three dates all centers are missing. For these dates the operational ECMWF analysis is used.

As the forecasts were evaluated on pressure levels, it was important to properly mask grid points on each level that fall below the model orography. This could either be done separately for each forecast step and model, or with a fixed mask. In this project we opted for a fixed mask. The mask is determined for each level by the following steps:

1) Determine the minimum value of the geopotential height over all models, dates, and lead times.
2) Interpolate the model orography from each model to 0.5° with a maximum interpolation method (the maximum value of the contributing grid points are used).
3) Determine the maximum orography among the contributing models and add 20% of the height to avoid near-surface effects.

For the precipitation verification, the NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) dataset is used (Hong et al. 2004). It is a high-frequency (half-hourly), high-resolution (0.1°) satellite observational dataset covering the global area between 60°S and 60°N. Pradhan et al. (2022) note that Asia (which we will highlight in the following sections) is the subject of the most IMERG evaluation studies on the continental and country scale. Given the tremendous uncertainty in global precipitation analyses, the results may be sensitive to the choice of validation dataset (Gehne et al. 2016), and further investigation using other high-resolution precipitation datasets [e.g., Multi-Source Weighted-Ensemble Precipitation (MSWEP); Beck et al. 2019] is warranted.

Initial conditions. All models in the comparison are initialized from ECMWF operational analyses based on IFS model cycle 45r1 (ECMWF 2018). The analysis fields are produced from the (mainly) strong-constraint 4D-Var data assimilation (Rabier et al. 2000) with a 6-h assimilation window (±3 h around the initialization time), while the first-guess forecast is provided from an analysis based on a 12-h window. To provide background-error statistics, a 25-member ensemble of 4D-Var assimilations is run with a lower horizontal resolution (Bonavita et al. 2012).

The analysis has a 9-km horizontal grid spacing and 137 vertical levels. In this project the atmospheric variables were interpolated to a regular 0.1° grid and distributed to all participating institutes (DWD, Météo-France, and Met Office accessed the data directly from the ECMWF archive). The number of variables used from the ECMWF analyses varies among the institutes (Table 3). The SST and sea ice were used from the Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) analysis Good et al. (2020) for all models except SHiELD. For land (soil + snow) all institutes used data from their own analysis, except for DWD that used the ECMWF analysis. We recognize that there is a risk that the inconsistencies between the initial boundary conditions (e.g., orography and land surface initial conditions) and the atmospheric initial conditions could impact medium-range forecast skill, especially for surface related variables (Boisserie et al. 2016). However, the implementation of the different surface models made it too difficult to ingest the ECMWF surface analysis into each model.

A complication in this experimental setup is the potential initialization adjustment/shock from starting a model with initial conditions produced by a different model. The initial shock could arise from differences due to interpolation of the initial conditions, differences in the lower boundary conditions (orography and land initial conditions) between models, and different model microphysics. For the microphysics, some institutes used parts of the available cloud water variables from the ECMWF analysis, while other institutes started with zero cloud water (0 in Table 3). The initialization shock could potentially act in a similar way to initial perturbations between the different model forecasts (Judd et al. 2008). These could then grow with time if they project onto growing dynamical modes.

To illustrate the initialization adjustment/shock, Fig. 1 shows the mean precipitation averaged between 40°N and 40°S for 24-h sliding windows with increasing lead times for all models and the GPM precipitation product (Hong et al. 2004). This diagnostic of spinup/-down of precipitation serves as an indication of the initial shock in the models. For this measure, the largest shock is seen for CMC-GEM, which resulted in very low precipitation in the beginning of the forecast. To test the sensitivity to the initialization of humidity, an additional experiment was run with the humidity from the native CMC-GEM analysis (CMC-GEM-Q). This resulted in a shock in the other direction with too much precipitation in the beginning of the forecast. The plot also includes forecasts from the native (own) analysis for UM, JMA-GSM, and for SHiELD based on GFS initial conditions. For these datasets a spinup period is still present, but not as strong as starting from the ECMWF analyses.

Table 3. Variables used from the ECMWF analysis for initializing each model. T = Temperature, U = zonal wind component, V = meridional wind component, Q = specific humidity, W = vertical velocity, CLWC = cloud liquid water content, CIWC = cloud ice water content, CRWC = cloud rainwater content, CSWC = cloud snow water content, CC = cloud fraction, PS = surface pressure.

Model	3D fields	Microphysics	2D fields
CMC-GEM	T, U, V, Q	0	PS, orography
ICON	T, U, V, Q, W	CLWC, CIWC	PS, skin temperature, soil and snow parameters
SHiELD	T, U, V, Q, W	CLWC, CIWC, CRWC, CSWC	PS, orography
JMA-GSM	T, U, V, Q	0	PS, orography
Arpege	T, U, V, Q	0	PS
UM	T, U, V, Q	CLWC, CIWC, CC	PS

The smallest magnitude shock is found for

ECMWF (even if the used model version is a newer version of IFS than used for the initial conditions), followed by ICON. In general, the adjustment time scale is around 3 days before the model reaches a relative equilibrium. But from this plot one can also notice that different models have different mean precipitation, something to be discussed more in the next section. The lead-time dependence of the 500-hPa temperature bias and possible initial shocks will be also discussed in the next section.

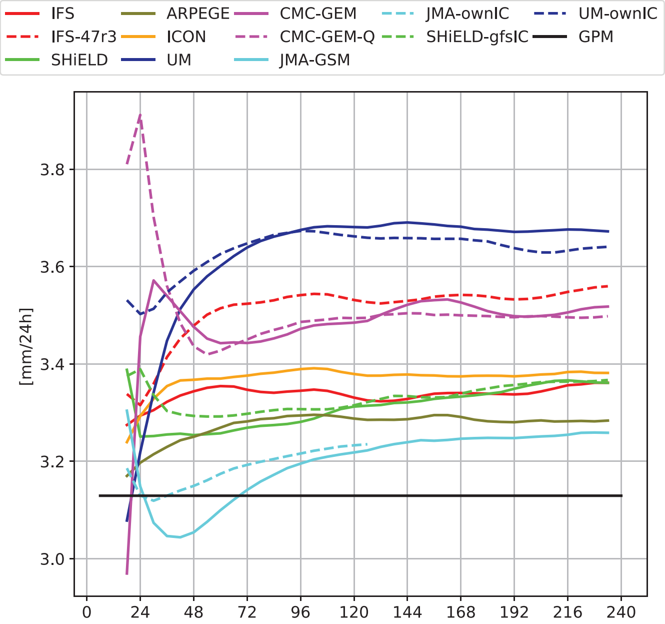

Fig. 1. 24-h mean precipitation for running-mean windows for 40°N–40°S, as a function of lead time (h).

Results

In this section we will present a selection of bias-corrected root-mean-square error/difference (RMSE/D), forecast standard deviation (hereafter referred to as "forecast activity"), and mean error (bias) statistics. The bias-corrected RMSE is calculated by subtracting the lead-time-dependent mean error of the full sample from each forecast before calculating the RMSE. Note that the bias-corrected RMSE is equivalent to the standard deviation of forecast errors. The forecast activity measures the standard deviation of the forecast minus the daily climatology from the ERA-Interim, and is important to monitor as reduced activity can artificially decrease RMSE (see chapter 12.A in ECMWF 2022).

Lead-time evolution of RMSE and bias. For each model we have calculated RMSE, bias and forecast activity for a set of parameters and levels. Figure 2 shows an example for the results for 500-hPa temperature (T500) over the Northern Hemisphere (N. Hemisphere, 20°–90°N) and the tropics (20°N–20°S). The results for the Southern Hemisphere (S. Hemisphere, 20°–90°S) will be discussed but are not shown. The forecasts are verified against the multicenter analysis (described above). The figures include the bias (Figs. 2a,b), the bias-corrected RMSE (Figs. 2c,d; thick lines), forecast activity (Figs. 2c,d; thin lines), and the difference in the bias-corrected RMSE between each model and the IFS-47r1 forecasts (Figs. 2e,f). In the difference plot, the results that are statistically different to IFS-47r1 (at the 95% level using the Student's *t* test) are marked with a dot. The scores have been calculated with 12-hourly time increments.

In these figures the verification for the subsequent IFS model (IFS-47r3, operational in autumn 2021) is included to give an example of the incremental change obtained for an upgrade of one model. We have also included the forecasts from JMA (up to day 6; JMA-ownIC) and the UM (UM-ownIC) initialized from their own analyses, and the SHiELD model initialized with the NCEP/GFS initial conditions (SHiELD-gfsIC), to illustrate the impact of the choice of initial conditions. Note that the NCEP/GFS initial conditions are from before the operational implementation of the FV3 dynamical core (same core as SHiELD) in GFS on 12 June 2019.

For all forecasts initialized from the ECMWF analysis, the bias starts at values close to each other. The error at step 0 for the ECMWF forecast indicates the difference (both in mean and RMSE) between the ECMWF analysis and the multicenter analysis. The initialization from the ECMWF analysis (based on a previous model version) may give the IFS forecasts an advantage due to a smaller initialization shock.

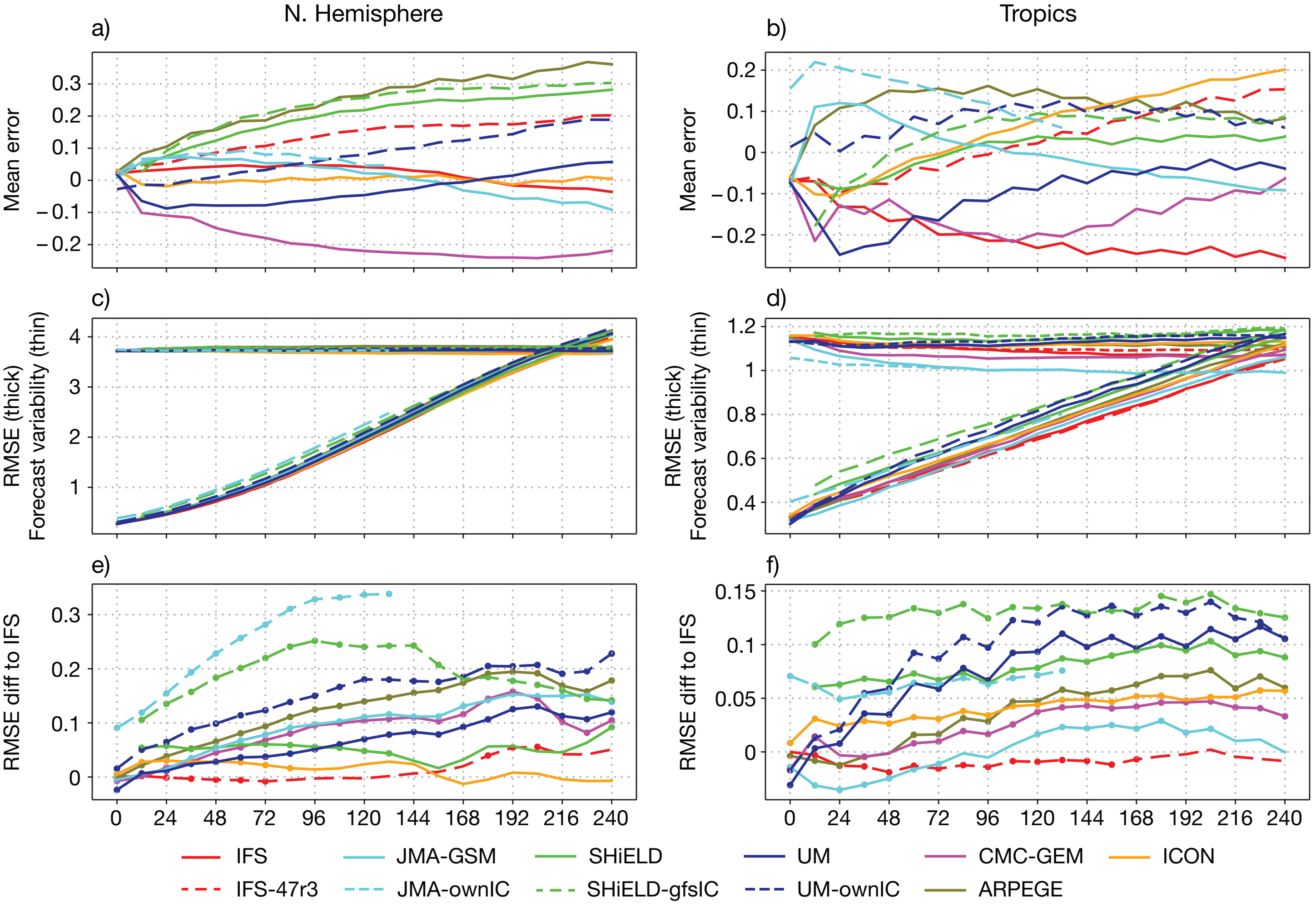

Fig. 2. Lead-time evolution for T500 of (a),(b) mean error; (c),(d) bias-corrected RMSE (thick lines) and forecast standard deviation (thin lines); and (e),(f) difference in bias-corrected RMSE to IFS. For the difference plot, the lead times with statistically significant results above the 95% level are marked with dots. (left) N. Hemisphere and (right) tropics.

For the bias, the models start to deviate over the first 12-h with UM and CMC quickly developing a negative bias. This could be a result of initialization shock when using an analysis from a different system, or a fast-developing bias in these models. The UM subsequently "warms up" after the initialization, but stays colder than the UM forecasts initialized from its own analysis for the full forecast range. For the N. Hemisphere and S. Hemisphere (not shown), Arpege, SHiELD, and IFS-47r3 develop a positive bias with increasing lead time. The warming in IFS-47r3 is not present for IFS-47r1, which has a neutral temperature bias for the N. Hemisphere. Conversely, there is a cold bias in the tropics for IFS-47r1, which is reduced in IFS-47r3. The vertical structure of these biases is discussed later in this section. For the tropics we find a rapidly developing positive bias for Arpege and JMA-GSM; however, JMA-GSM later "cools down." For longer lead times, ICON drifts toward warmer conditions in the tropics at 500 hPa.

For the lead-time-dependent RMSE, as seen in Fig. 2 for T500 over N. Hemisphere, the errors grow with increasing lead time (as expected) and do not reach saturation by day 10. From day 1 to day 3 the IFS forecast for T500 in the N. Hemisphere is significantly better than all other models, while ICON obtains similar RMSE for longer lead times. For the RMSE in both the N. Hemisphere and tropics, we find the largest errors for the three forecasts with other initial conditions than ECMWF (JMA and UM with own analysis and SHiELD initialized from NCEP/GFS analyses). Conversely, when these models are initialized from ECMWF analyses, the errors are much lower. The relatively small difference between the models initialized from ECMWF analysis, compared to the three forecasts with different analyses, we see as a sign

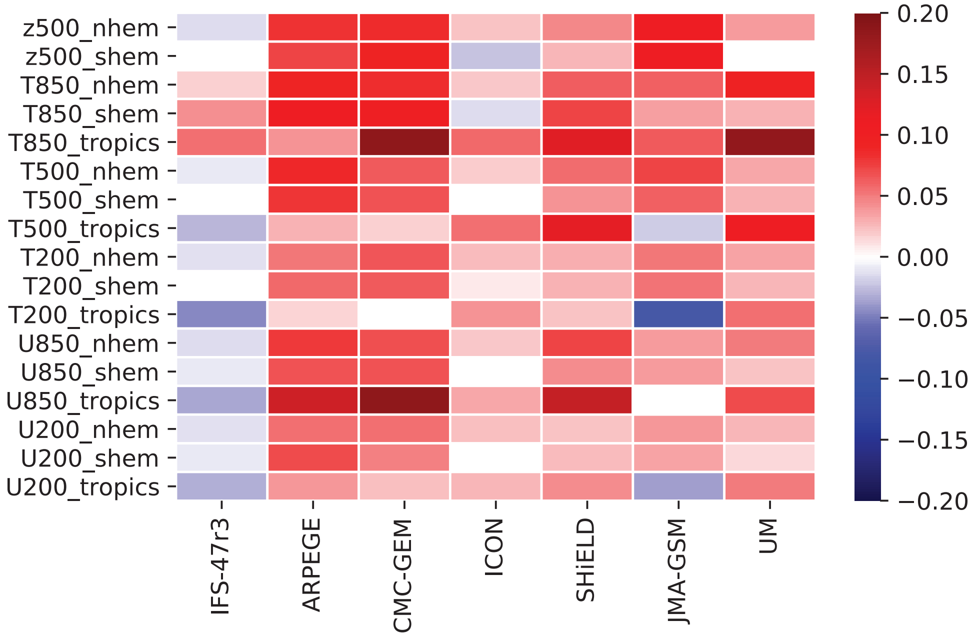

Fig. 3. Scorecard of normalized, bias-corrected RMSE difference to IFS, step 72 h, for different variables, levels, and regions.

that all models are capable of producing high-quality forecasts and that the initial conditions play a significant role.

For the S. Hemisphere (not shown) the results are similar to the N. Hemisphere but with a larger difference between SHiELD initialized from NCEP/GFS analyses and all models initialized from ECMWF analyses. For the S. Hemisphere the lowest RMSE is found for ICON, but the result only passes the 95% significance around day 4.

For the T500 RMSE in the tropics, we find the lowest values for JMA-GSM and IFS-47r3. However, for JMA-GSM we simultaneously find a too low forecast activity that could at least partly explain the lower RMSE. The decrease in activity is found to happen over the first 2 forecast days and later converges to JMA-ownIC. A similar decrease in forecast activity is also found for CMC-GEM that could contribute to the relatively low RMSE. The RMSEs from ICON and Arpege are comparable with those of IFS-47r1. One can note that these three models share the same convection scheme.

Figure 3 summarizes the 3-day normalized difference in bias-corrected RMSE to the IFS-47r1 forecast for the N. Hemisphere, S. Hemisphere, and the tropics regions for a range of parameters. All nonsignificant (using 95% significance level from the Student's *t* test) differences are masked white in this table. Except for the T850, we can see the improvement with the IFS-47r3 model version compared to IFS-47r1. Among the models, ICON has the most similar scores to the IFS for most of the parameters and regions, and is better than both versions of IFS for T850 and Z500 in the S. Hemisphere. We can also see that JMA-GSM has lower RMSEs than all other models for T200 and U200 in the tropics. This could be an artifact of the reduced anomalies (see Fig. 2 for T500) in the JMA-GSM forecasts that can favor the RMSE metric.

Temperature bias in the troposphere. Figure 4 shows the zonal mean of the day 3 forecast temperature bias, based on data on 850-, 700-, 500-, 300-, and 200-hPa levels. The first pattern to notice is that all models have a cold bias in the upper troposphere at all latitudes, which is most pronounced outside the tropics. This could be a result of the ECMWF initial conditions being too moist in the lower stratosphere, which causes strong radiative cooling (Shepherd et al. 2018).

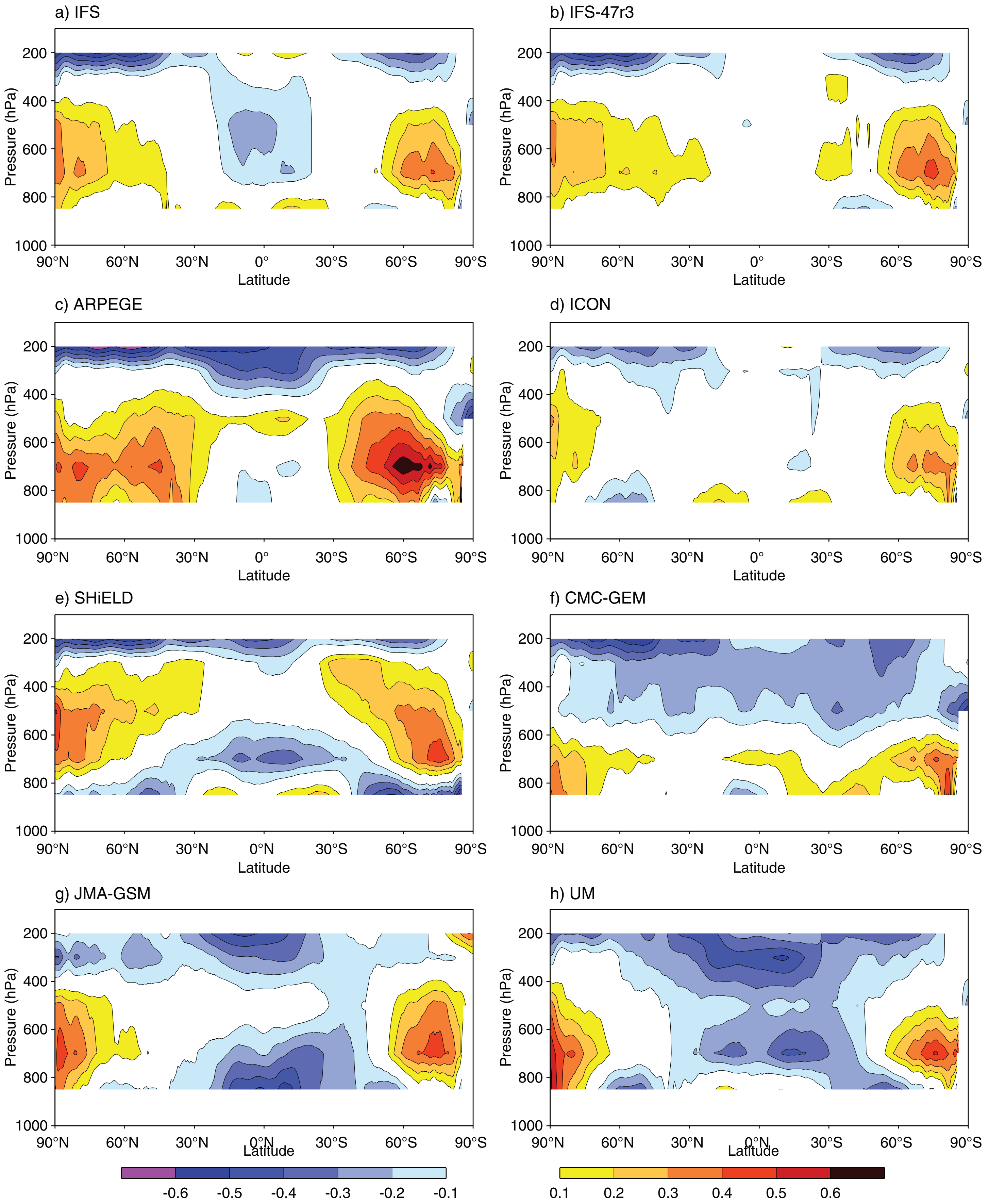

Fig. 4. Vertical cross section of zonal mean temperature bias, step 72 h, verified against TIGGE multianalysis.

All model simulations have a warm bias at high latitudes in the middle troposphere. For the tropics, the vertical structure of the bias differs a lot between the models, where UM is the coldest.

Comparing the bias pattern at 850 hPa (Fig. 5), we see large regional differences. SHiELD simulations have a cold bias over the northern Atlantic and northern Pacific while ICON is too cold over the landmasses. JMA-GSM has a cold bias over the lower latitude oceans (south of 30°N) and UM has a mix of cold and warm regional biases. Arpege has the strongest warm bias over oceans outside the tropics.

Over the subtropical regions, which are dominated by strong inversions around the 850-hPa level, the IFS, CMC-GEM, and SHiELD have a warm bias, whereas the UM is too cold, and ICON and Arpege show small biases. Torn and Davis (2012) compared the Tiedke (used here in IFS, ICON, Arpege) with the Kain–Fritsch (used here in CMC-GEM) convection scheme over tropical oceans and found the former performs better. Even if our results agree, one can note the large impact by the recent changes in existing schemes in IFS-47r3.

For the lower troposphere over the S. Hemisphere, JMA-GSM, UM, and SHiELD have the strongest cold biases. Again it is the lower latitudes that contribute most to the bias in JMA-GSM, while UM and SHiELD have stronger biases over the storm track region between 40° and 60°S.

Precipitation bias over Southeast Asia. Southeast Asia provides a complex region in terms of precipitation with the mixture of land and warm sea around the Maritime Continent, monsoon-driven variability over the continental landmasses, the mei-yu front over China, and finally orographic enhancement upwind of the Himalayas. We therefore chose to exemplify the precipitation biases in this region (Fig. 6). Here, the bias is averaged over the full 10-day period in each forecast and verified against the GPM precipitation product (Hong et al. 2004).

The most common bias among the models is the dry bias over the eastern Indian Ocean off the coast of Sumatra, which is present in all models but ICON (wet bias). The strongest bias is found in SHiELD and Arpege, where the dry bias extends farther west over the Indian Ocean. The underestimation is related to a lack of 10-day forecasts with high precipitation rates (not shown).

For ICON we find a strong contrast in the precipitation bias between the land and ocean. Precipitation amounts are too high over the sea and too low over the land compared to GPM. Several of the other models also simulate a wet bias over sea around the Maritime Continent. This difference between ICON and IFS warrants further analysis as they both use the same convection scheme.

Another common bias among the models is excess precipitation over southeastern China, which appears in all models except ICON. This bias is discussed in Lavers et al. (2021) for IFS, where the forecasts are compared to rain gauge observations. We found too many 10-day forecasts with high precipitation rates in the region (not shown). The UM and IFS also simulate excessive precipitation along the southeastern edge of the Himalaya Mountains.

Multimodel ensemble spread. Historically, forecasters have used deterministic forecasts from multiple NWP centers to either subjectively assess the forecast uncertainty or create a multimodel ensemble (Ziehmann 2000). The uncertainty in such an ensemble would be a result of the differences between analyses from each NWP center together with the difference due to model formulations. Though none of these components have been developed to represent the true uncertainty, the methodology is popular.

In operational ensemble forecasting systems from a single NWP center, dedicated schemes are used to account for the analysis and model uncertainties. For the latter, one commonly

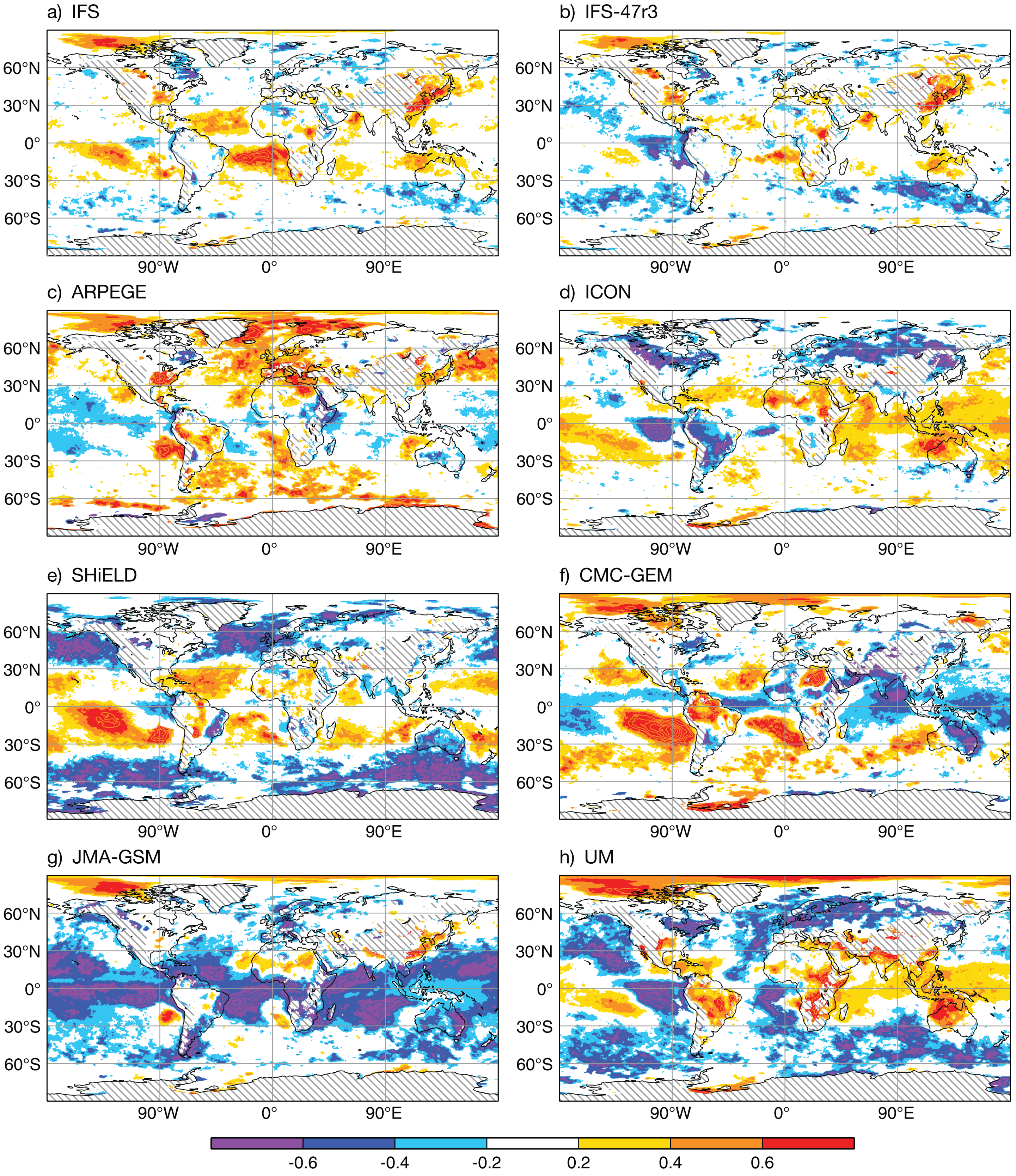

Fig. 5. 850-hPa temperature bias, step 72 h, verified against TIGGE multianalysis. Hatches mark the orography mask for 850 hPa.

used scheme is the stochastically perturbed physics tendencies (SPPT) scheme (Buizza et al. 1999; Leutbecher et al. 2017). The SPPT method perturbs the total tendency of the physical parameterization schemes and aims to only target the random part of the model error. Another example, closer to the multimodel approach, is to use different combinations of physical

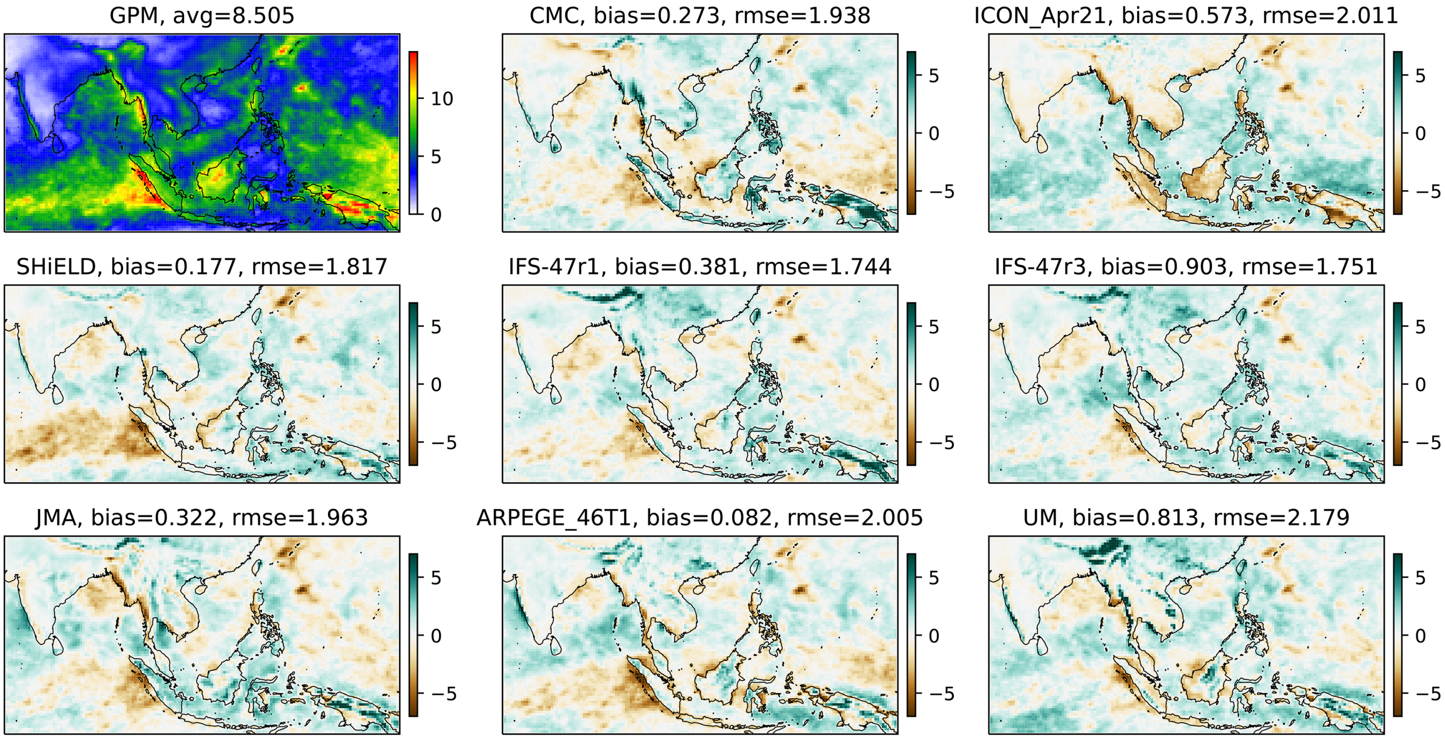
GPM, avg=8.505
10
5
0
CMC, bias=0.273, rmse=1.938
ICON_Apr21, bias=0.573, rmse=2.011
SHiELD, bias=0.177, rmse=1.817
IFS-47r1, bias=0.381, rmse=1.744
IFS-47r3, bias=0.903, rmse=1.751
JMA, bias=0.322, rmse=1.963
ARPEGE_46T1, bias=0.082, rmse=2.005
UM, bias=0.813, rmse=2.179
5
0
−5

Fig. 6. (top left) Mean precipitation from GPM and (remaining panels) precipitation bias for each model averaged over 10-day forecasts against GPM.

parameterization schemes such as in Descamps et al. (2015). For an overview of other schemes, see Leutbecher et al. (2017).

In this section we evaluate the multimodel ensemble standard deviation (hereafter referred to as ensemble spread) based on seven models in DIMOSIC (IFS-47r3 excluded). As all models are initialized from the same initial conditions, we expect the forecast differences to arise mainly from differences in the model, bearing in mind that the initial adjustment can also create growing perturbations. The mean differences between the models are removed before computing the multimodel ensemble spread. This partly addresses the effect of the initial adjustment. The multimodel ensemble is compared with the ensemble spread from a 10-member ECMWF ensemble with the SPPT scheme but without initial perturbations.

In Fig. 7 we compare the T500 ensemble spread for the operational ECMWF ensemble, the DIMOSIC multimodel ensemble, and a 10-member ECMWF ensemble with the SPPT scheme but without initial perturbations. The ensemble spread has been scaled to compensate for the finite ensemble size (Leutbecher 2009). The plot also includes the ensemble-mean RMSE for the ECMWF ensemble, verified against the ECMWF analysis. As the DIMOSIC ensemble and the SPPT ensemble excludes initial uncertainties, we do not expect these to simulate the full forecast uncertainty.

For the N. Hemisphere, the ECMWF ensemble is reliable as the ensemble spread matches closely with the ensemble-mean error (Leutbecher and Palmer 2008), while the ensemble spread is less for the two other ensembles as expected. The ensemble spread from the DIMOSIC ensemble grows rapidly during the first 24 h, which could be due to fast adjustments in each model and/or due to fast-growing uncertainties on the convective scale, e.g., discussed in Zhang et al. (2019). In 3-day forecasts the spread is slightly higher in the DIMOSIC ensemble compared to the SPPT ensemble. Inspecting maps of the spread for this lead time (Fig. 8) one finds that such a difference is present in the Arctic and over the landmasses, while in the storm tracks the spread is similar in the two ensembles. In the tropics, the ECMWF ensemble

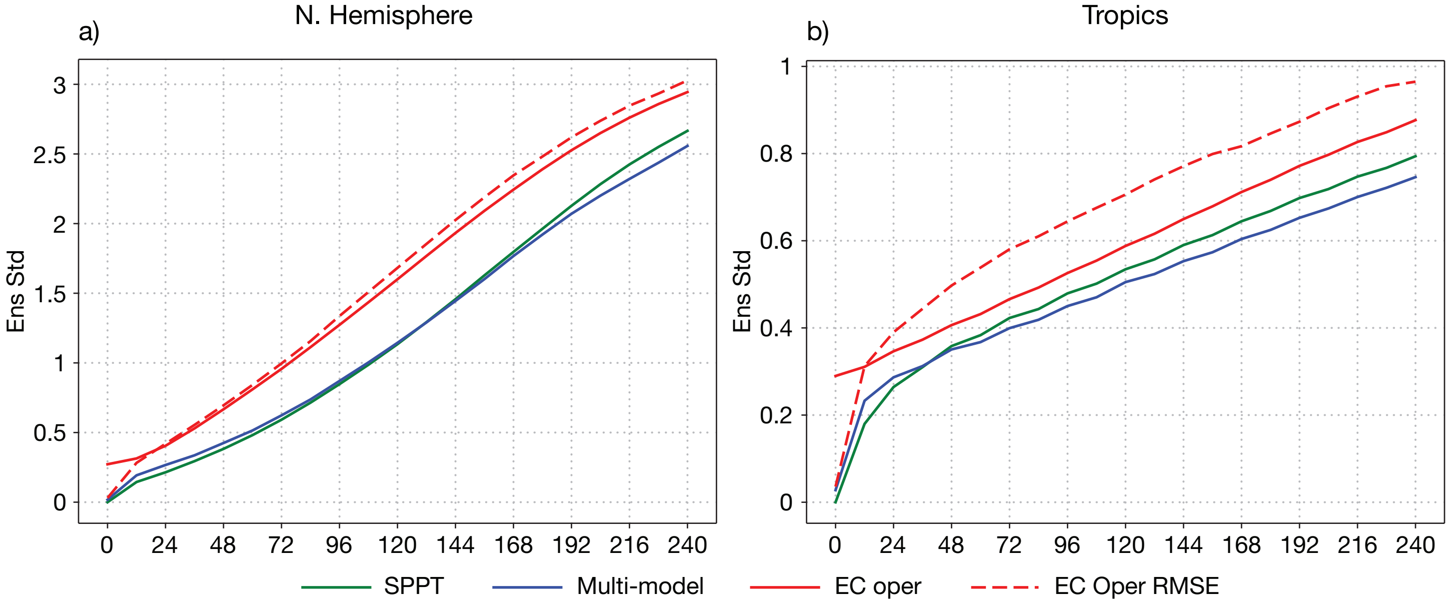

Fig. 7. Lead-time evolution of ensemble spread (solid) in ECMWF operational ensemble (red), DIMOSIC multimodel ensemble (blue), and SPPT ensemble (green). Ensemble-mean RMSE of ECMWF's operational ensemble (red dashed). (a) N. Hemisphere and (b) tropics.

spread is less than the ensemble-mean RMSE for T500. However, here the ensemble-mean RMSE has not been bias corrected. Furthermore, it is important to take analysis uncertainty into account when judging ensemble reliability (see Lang et al. (2021) for a discussion of ensemble verification sensitivities). Comparing the DIMOSIC ensemble with the SPPT ensemble we find a larger spread for the SPPT ensemble at 3-day lead time. Inspecting Fig. 8, we find the main difference over the ITCZ, where we expect strong model tendencies to be perturbed by the SPPT scheme.

In this section we have given examples of diagnostics of the ensemble spread in the DIMOSIC multimodel ensemble and how it can be compared to a dedicated model uncertainty scheme. Overall, the multimodel ensemble spread is similar to the SPPT ensemble, which is an interesting finding as the multimodel spread is not specifically targeted to represent model uncertainty. Future work will include a more comprehensive comparison of the multimodel ensemble and dedicated model uncertainty schemes.

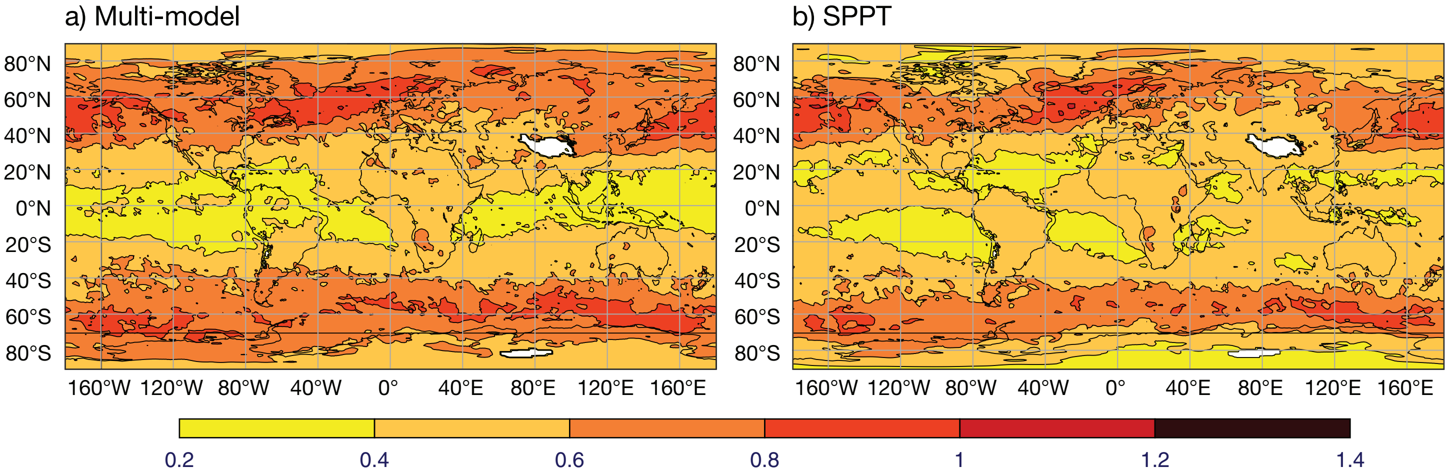

Fig. 8. Ensemble spread for temperature at 500 hPa at day 3 for (a) DIMOSIC multimodel ensemble spread and (b) SPPT ensemble.

Table 4. Average RMSD to the other models at step 72 h, averaged over the full globe.

Model	T850	T500	T200
CMC-GEM	1.06	0.78	0.87
ICON	0.94	0.75	0.85
IFS	0.93	0.72	0.82
SHiELD	1.05	0.83	0.92
JMA-GSM	0.99	0.78	0.87
Arpege	1.00	0.80	0.87
UM	1.01	0.77	0.88

Forecast differences between models. In this section we explore if some models produce forecasts that are more similar to each other. This is done by calculating the root-mean-square difference (RMSD) between pairs of models after applying a bias correction (same as used above), for specific regions. We also calculate the RMSE (RMSD to the multianalysis), with the same procedure for each model.

Table 4 shows RMSD for each model averaged over all pairs with other models (excluding IFS-47r3 and the multianalysis), for 3-day forecasts and averaged over the full globe. For all three temperature levels the lowest average RMSD is found for pairs involving IFS. The largest difference is found for CMC-GEM on 850 hPa and for SHiELD for 500 and 200 hPa, indicating that these two models are most different relative to the rest of the models.

Figure 9 shows the RMSD between pairs of models for T500 at day 3. In the first column/row, the RMSD to the verifying analysis (error) is provided. Above the diagonal presents the results for N. Hemisphere and below for the tropics. The figures also include the result for the IFS-47r3 to show the difference obtained from two versions of the same model.

The models that are closest to each other in the N. Hemisphere are IFS and ICON, which are also the models that share several parameterization schemes (see Table 2) and have the lowest RMSE found in Fig. 2. But even if the models are close, the RMSD is much higher than between the two IFS versions. For the N. Hemisphere, the UM is also relatively close to both IFS and ICON.

For T500 in the tropics, one can first note that the bias-corrected error is similar or sometimes lower than the differences between the models. This result could be explained by the smoothing of the verifying analysis by using a multianalysis average. For the model differences,

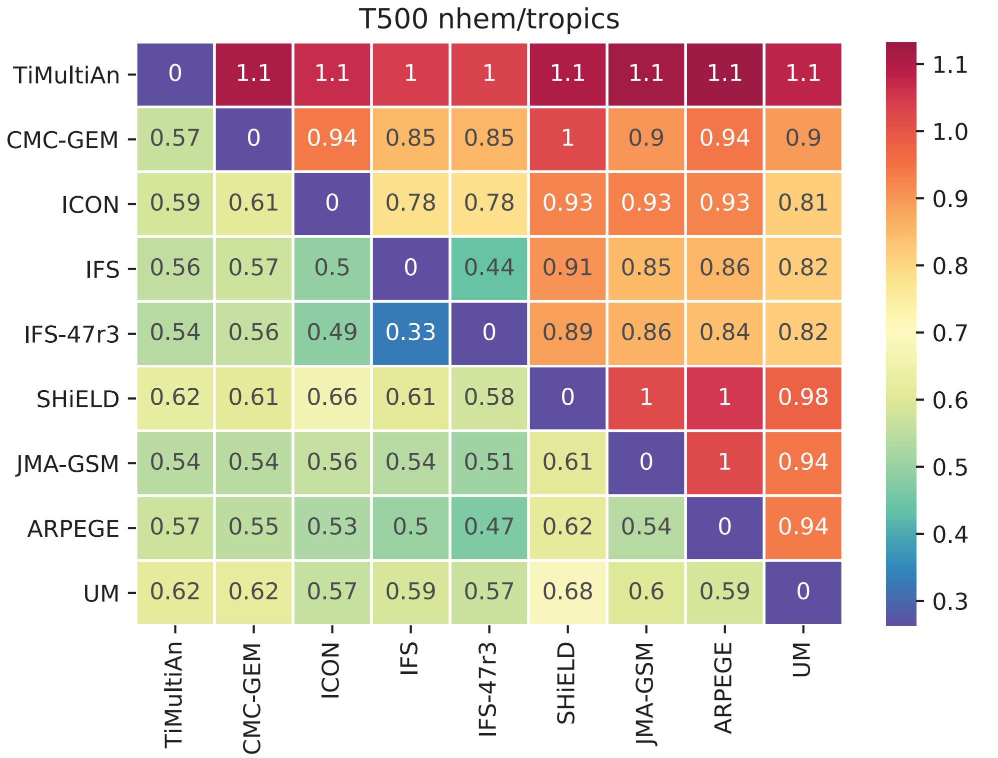

Fig. 9. Model pair RMS differences for T500, step 72 for N. Hemisphere (top-right triangle) and tropics (bottom-left triangle).

ICON, IFS, and Arpege are the models closest to each other. These three models use a similar convection parameterization (based on the Tiedtke–Bechtold scheme), which could explain the similarities.

Discussion and conclusions

In this paper we have presented examples of results from the model intercomparison project DIMOSIC, where different global models have been run from the same initial conditions. We have presented results for the RMSE of the forecasts, model biases and standard deviation of a multimodel ensemble, and differences between pairs of models.

The initial conditions came from the ECMWF operational data assimilation. The choice was based on most of the global modeling centers having some experience from initializing from ECMWF operational or reanalyses (e.g., ERA5). However, some models experienced a large initialization shock in terms of average precipitation, something that is to be investigated further.

For the *quality of the forecasts*, for T500 over the N. Hemisphere, all models initialized with ECMWF analysis produced forecasts with RMSE much more similar to each other than forecasts initialized from their own analyses. It shows that all models in this study are capable of producing high-quality forecasts given a high-quality analysis, and that the initial shock seems to not cause a significant harm to the forecast quality. This result is in line with Magnusson et al. (2019), where a similar improvement was seen for the GFDL/FVGFS model (an earlier version of the SHiELD model), by using the ECMWF analyses.

Among the models the IFS showed the lowest RMSE for most parameters and regions. The result could be partly because we expected the lowest initialization shock from IFS. The second-best-performing model was for most parameters ICON, and that model had the lowest RMSE for Z500 in the medium-range over the S. Hemisphere.

For the *difference in model biases*, we found a large variety, both in terms of temperature errors and precipitation. The DIMOSIC dataset provides a convenient way to investigate particular biases with the effect of different mean states of the initial conditions removed. Except for the examples discussed in this article, the DIMOSIC dataset has for example been used at ECMWF to understand biases in 700-hPa temperature during boreal summer (Magnusson et al. 2022).

A large *diversity between forecasts* was found when we compared the RMSD between the models, much higher than between two versions of IFS. ICON and IFS came out as the most similar in terms of RMSD between the models. This is not a surprise as the ICON model shares parts of its physical parameterizations with IFS. At the same time, the pair of ICON and SHiELD showed relatively large RMSD. However, it is difficult to point out a single model component that has the strongest impact on the forecast differences. One has to bear in mind that interaction between different model components and configurations of each component play a significant role as well. Nevertheless, the information of regional differences in forecast skill and biases will give guidance toward the responsible model process.

For the *multimodel ensemble spread*, we showed examples of the horizontal distribution of the uncertainty and compared the results with an ensemble using stochastically perturbed parameterization tendencies (SPPT; Leutbecher et al. 2017). The two different approaches to capture model uncertainty showed similarities in terms of average ensemble spread.

In the future we plan to investigate the flow dependency of the model uncertainty and differences between model formulations, for example, by investigating the relation to warm conveyor belts as discussed in Rodwell and Wernli (2022, manuscript submitted to *Wea. Climate Dyn.*). To facilitate such a study, we plan to make use of the Eulerian detection of warm conveyor belts (Quinting and Grams 2022).

The DIMOSIC project has provided a benchmark dataset that has already helped the different NWP centers with model development. For example, it has been greatly beneficial during the development of the Navy Environmental Prediction System Utilizing a Nonhydrostatic Engine (NEPTUNE) global model at NRL to identify systematic biases and also contributed to the identification of errors in other models. The dataset has also been used at KMA to benchmark the newly developed Korean Integrated Model (KIM).

In this article we have focused on the forecasts of field variables (temperature, wind, geopotential, and precipitation). Ongoing work in the project will evaluate tropical and extratropical cyclones in the different forecasts and the results will be presented separately. In this article we have based all results on evaluation spanning over a full calendar year, with forecasts initialized every third day. There is a scope to evaluate the different seasons separately.

In summary, all contributing models produce forecasts with high skill when initialized from the ECMWF initial conditions, compared to the forecast initialized from other analyses. In a deterministic sense this suggests that the initial conditions are a stronger factor to create the diversity in forecast skill between NWP centers than the model formulations. However, the forecast difference is still large between the models, as well as the multimodel ensemble spread. This leads us to the conclusion that the parallel developments at different NWP centers provide diversity among models. This will in the long term hopefully enhance overall progress in the field of NWP.

For the future, DIMOSIC provides a framework for comparisons of global models that gives model developers an additional tool to use to put their results in perspective to other models and also gives further insights about the diversity of the existing global models for forecasters and end users. With new model developments, new runs with the protocol will be evaluated. The DIMOSIC dataset will be available for the research community for further explorations.

Acknowledgments. We thank the three reviewers that helped us to improve this article. We would like to acknowledge David Richardson, Florian Pappenberger, and Richard Forbes for their input to this article. A special thanks to Anabel Bowen for helping to prepare the graphics for this article. James Doyle and Alex Reinecke acknowledge the support of the Office of Naval Research PE 0602435N. Duncan Ackerley was supported by the Met Office Hadley Centre Climate Programme funded by BEIS and Defra.

Data availability statement. All model data used in this paper can be requested from the corresponding author.

References

Bauer, P., L. Magnusson, J. Thepaut, and T. M. Hamill, 2016: Aspects of ECMWF model performance in polar areas. *Quart. J. Roy. Meteor. Soc.*, **142**, 583–596, https://doi.org/10.1002/qj.2449.

Bechtold, P., M. Köhler, T. Jung, F. Doblas-Reyes, M. Leutbecher, M. J. Rodwell, F. Vitart, and G. Balsamo, 2008: Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales. *Quart. J. Roy. Meteor. Soc.*, **134**, 1337–1351, https://doi.org/10.1002/qj.289.

——, R. Forbes, I. Sandu, S. Lang, and M. Ahlgrimm, 2020: A major moist physics upgrade for the IFS. *ECMWF Newsletter*, No. 164, ECMWF, Reading, United Kingdom, 24–32, https://www.ecmwf.int/en/newsletter/164/meteorology/major-moist-physics-upgrade-ifs.

Beck, H. E., E. F. Wood, M. Pan, C. K. Fisher, D. G. Miralles, A. I. J. M. van Dijk, T. R. McVicar, and R. F. Adler, 2019: MSWEP V2 global 3-hourly 0.1 precipitation: Methodology and quantitative assessment. *Bull. Amer. Meteor. Soc.*, **100**, 473–500, https://doi.org/10.1175/BAMS-D-17-0138.1.

Bengtsson, L., and Coauthors, 2019: Convectively coupled equatorial wave simulations using the ECMWF IFS and the NOAA GFS cumulus convection schemes in the NOAA GFS model. *Mon. Wea. Rev.*, **147**, 4005–4025, https://doi.org/10.1175/MWR-D-19-0195.1.

Boisserie, M., B. Decharme, L. Descamps, and P. Arbogast, 2016: Land surface initialization strategy for a global reforecast dataset. *Quart. J. Roy. Meteor. Soc.*, **142**, 880–888, https://doi.org/10.1002/qj.2688.

Bonavita, M., L. Isaksen, and E. Holm, 2012: On the use of EDA background error variances in the ECMWF 4D-Var. *Quart. J. Roy. Meteor. Soc.*, **138**, 1540–1559, https://doi.org/10.1002/qj.1899.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, https://doi.org/10.1002/qj.49712556006.

Chen, J., and Coauthors, 2019: Advancements in hurricane prediction with NOAA's next-generation forecast system. *Geophys. Res. Lett.*, **46**, 4495–4501, https://doi.org/10.1029/2019GL082410.

Dalcher, A., and E. Kalnay, 1987: Error growth and predictability in operational ECMWF forecasts. *Tellus*, **39A**, 474–491, https://doi.org/10.1111/j.1600-0870.1987.tb00322.x.

Descamps, L., C. Labadie, A. Joly, E. Bazile, P. Arbogast, and P. Cébron, 2015: PEARP, the Meteo-France short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **141**, 1671–1685, https://doi.org/10.1002/qj.2469.

DWD, 2022: ICON: Icosahedral nonhydrostatic weather and climate model. https://code.mpimet.mpg.de/projects/iconpublic/wiki/Documentation.

ECMWF, 2018: IFS documentation CY45r1. ECMWF Tech. Rep., https://www.ecmwf.int/en/publications/ifs-documentation.

——, 2020: IFS documentation CY47r1. ECMWF Tech. Rep., https://www.ecmwf.int/en/publications/ifs-documentation.

——, 2022: ECMWF forecast user guide. ECMWF Tech. Rep., https://confluence.ecmwf.int/display/FUG/Forecast+User+Guide.

Gehne, M., T. M. Hamill, G. N. Kiladis, and K. E. Trenberth, 2016: Comparison of global precipitation estimates across a range of temporal and spatial scales. *J. Climate*, **29**, 7773–7795, https://doi.org/10.1175/JCLI-D-15-0618.1.

Girard, C., and Coauthors, 2014: Staggered vertical discretization of the Canadian Environmental Multiscale (GEM) model using a coordinate of the log-hydrostatic-pressure type. *Mon. Wea. Rev.*, **142**, 1183–1196, https://doi.org/10.1175/MWR-D-13-00255.1.

Good, S., and Coauthors, 2020: The current configuration of the OSTIA system for operational production of foundation sea surface temperature and ice concentration analyses. *Remote Sens.*, **12**, 720, https://doi.org/10.3390/rs12040720.

Harris, L., and Coauthors, 2020: GFDL SHiELD: A unified system for weather-to-seasonal prediction. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002223, https://doi.org/10.1029/2020MS002223.

Hong, Y., K.-L. Hsu, S. Sorooshian, and X. Gao, 2004: Precipitation Estimation from Remotely Sensed Imagery Using an Artificial Neural Network Cloud Classification System. *J. Appl. Meteor.*, **43**, 1834–1853, https://doi.org/10.1175/JAM2173.1.

JMA, 2019: Outline of the operational numerical weather prediction at the Japan Meteorological Agency. Japan Meteorological Agency, http://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline2019-nwp/index.htm.

Judd, K., C. A. Reynolds, T. E. Rosmond, and L. A. Smith, 2008: The geometry of model error. *J. Atmos. Sci.*, **65**, 1749–1772, https://doi.org/10.1175/2007JAS2327.1.

Judt, F., and Coauthors, 2021: Tropical cyclones in global storm-resolving models. *J. Meteor. Soc. Japan*, **99**, 579–602, https://doi.org/10.2151/jmsj.2021-029.

Lang, S. T. K., S. Lock, M. Leutbecher, P. Bechtold, and R. M. Forbes, 2021: Revision of the stochastically perturbed parametrisations model uncertainty scheme in the integrated forecasting system. *Quart. J. Roy. Meteor. Soc.*, **147**, 1364–1381, https://doi.org/10.1002/qj.3978.

Lavers, D. A., S. Harrigan, and C. Prudhomme, 2021: Precipitation biases in the ECMWF integrated forecasting system. *J. Hydrometeor.*, **22**, 1187–1198, https://doi.org/10.1175/JHM-D-20-0308.1.

Leutbecher, M., 2009: Diagnosis of ensemble forecasting systems. *Annual Seminar: Diagnosis of Forecasting and Data Assimilation Systems*, Reading, United Kingdom, ECMWF, 32 pp., https://www.ecmwf.int/sites/default/files/elibrary/2010/10725-diagnosis-ensemble-forecasting-systems.pdf.

——, and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014.

——, and Coauthors, 2017: Stochastic representations of model uncertainties at ECMWF: State of the art and future vision. *Quart. J. Roy. Meteor. Soc.*, **143**, 2315–2339, https://doi.org/10.1002/qj.3094.

Lott, F., and M. J. Miller, 1997: A new subgrid-scale orographic drag parametrization: Its formulation and testing. *Quart. J. Roy. Meteor. Soc.*, **123**, 101–127, https://doi.org/10.1002/qj.49712353704.

Magnusson, L., J. Chen, S. Lin, L. Zhou, and X. Chen, 2019: Dependence on initial conditions versus model formulations for medium-range forecast error variations. *Quart. J. Roy. Meteor. Soc.*, **145**, 2085–2100, https://doi.org/10.1002/qj.3545.

——, M. Alonso-Balmaseda, M. Dahoui, R. Forbes, T. Haiden, D. Lavers, I. Sandu, and S. Tietsche, 2022: Summary of the UGROW subproject on tropospheric temperature bias during JJA over the northern hemisphere. EWMWF Tech. Memo. 891, 17 pp., https://www.ecmwf.int/node/20356.

McTaggart-Cowan, R., and Coauthors, 2019: Modernization of atmospheric physics parameterization in Canadian NWP. *J. Adv. Model. Earth Syst.*, **11**, 3593–3635, https://doi.org/10.1029/2019MS001781.

Mengaldo, G., A. Wyszogrodzki, M. Diamantakis, S.-J. Lock, F. X. Giraldo, and N. P. Wedi, 2019: Current and emerging time-integration strategies in global numerical weather and climate prediction. *Arch. Comput. Methods Eng.*, **26**, 663–684, https://doi.org/10.1007/s11831-018-9261-8.

Mogensen, K. S., L. Magnusson, and J.-R. Bidlot, 2017: Tropical cyclone sensitivity to ocean coupling in the ECMWF coupled model: Tropical cyclone sensitivity. *J. Geophys. Res. Oceans*, **122**, 4392–4412, https://doi.org/10.1002/2017JC012753.

Pollard, R. T., P. B. Rhines, and R. O. R. Y. Thompson, 1973: The deepening of the wind-mixed layer. *Geophys. Fluid Dyn.*, **4**, 381–404, https://doi.org/10.1080/03091927208236105.

Pradhan, R. K., and Coauthors, 2022: Review of GPM IMERG performance: A global perspective. *Remote Sens. Environ.*, **268**, 112754, https://doi.org/10.1016/j.rse.2021.112754.

Quinting, J. F., and C. M. Grams, 2022: EuLerian Identification of ascending AirStreams (ELIAS 2.0) in numerical weather prediction and climate models - Part 1: Development of deep learning model. *Geosci. Model Dev.*, **15**, 715–730, https://doi.org/10.5194/gmd-15-715-2022.

Rabier, F., H. Järvinen, E. Klinker, J.-F. Mahfouf, and A. Simmons, 2000: The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Quart. J. Roy. Meteor. Soc.*, **126**, 1143–1170, https://doi.org/10.1002/qj.49712656415.

Roehrig, R., and Coauthors, 2020: The CNRM global atmosphere model ARPEGE-Climat 6.3: Description and evaluation. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002075, https://doi.org/10.1029/2020MS002075.

Shepherd, T. G., I. Polichtchouk, R. Hogan, and A. Simmons, 2018: Report on stratosphere task force. ECMWF Tech. Memo. 824, https://www.ecmwf.int/node/18259.

Stevens, B., and Coauthors, 2019: DYAMOND: The DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains. *Prog. Earth Planet. Sci.*, **6**, 61, https://doi.org/10.1186/s40645-019-0304-z.

Swinbank, R., and Coauthors, 2016: The TIGGE project and its achievements. *Bull. Amer. Meteor. Soc.*, **97**, 49–67, https://doi.org/10.1175/BAMS-D-13-00191.1.

Tiedtke, M., 1993: Representation of clouds in large-scale models. *Mon. Wea. Rev.*, **121**, 3040–3061, https://doi.org/10.1175/1520-0493(1993)121<3040:ROCILS>2.0.CO;2.

Torn, R. D., and C. A. Davis, 2012: The influence of shallow convection on tropical cyclone track forecasts. *Mon. Wea. Rev.*, **140**, 2188–2197, https://doi.org/10.1175/MWR-D-11-00246.1.

van Niekerk, A., and Coauthors, 2020: Constraining ORographic Drag Effects (COORDE): A model comparison of resolved and parametrized orographic drag. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002160, https://doi.org/10.1029/2020MS002160.

Walters, D., and Coauthors, 2019: The Met Office unified model global atmosphere 7.0/7.1 and JULES global land 7.0 configurations. *Geosci. Model Dev.*, **12**, 1909–1963, https://doi.org/10.5194/gmd-12-1909-2019.

Williams, K. D., and Coauthors, 2013: The Transpose-AMIP II experiment and its application to the understanding of Southern Ocean cloud biases in climate models. *J. Climate*, **26**, 3258–3274, https://doi.org/10.1175/JCLI-D-12-00429.1.

Zeng, X., and A. Beljaars, 2005: A prognostic scheme of sea surface skin temperature for modeling and data assimilation: Sea surface skin temperature scheme. *Geophys. Res. Lett.*, **32**, L14605, https://doi.org/10.1029/2005GL023030.

Zhang, F., Y. Q. Sun, L. Magnusson, R. Buizza, S.-J. Lin, J.-H. Chen, and K. Emanuel, 2019: What is the predictability limit of midlatitude weather? *J. Atmos. Sci.*, **76**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.

Ziehmann, C., 2000: Comparison of a single-model EPS with a multimodel ensemble consisting of a few operational models. *Tellus*, **52A**, 280–299, https://doi.org/10.3402/tellusa.v52i3.12266.