Key Points:
- A climate forecast system is developed on the basis of a coupled climate model
- The comparison of the first to the second version shows improvements in North Atlantic Oscillation forecasts. Forecast skill in tropical regions is not improved

The German Climate Forecast System: GCFS

Kristina Fröhlich[1], **Mikhail Dobrynin[1,2]**, **Katharina Isensee[1]**, **Claudia Gessner[3,4]**, **Andreas Paxian[1]**, **Holger Pohlmann[1,5]**, **Helmuth Haak[5]**, **Sebastian Brune[2]**, **Barbara Früh[1]**, and **Johanna Baehr[2]**

[1]Deutscher Wetterdienst, Offenbach, Germany, [2]CEN, Universität Hamburg, Hamburg, Germany, [3]Goethe Universität Frankfurt, Frankfurt, Germany, [4]Now at ETH, Zurich, Switzerland, [5]Max Planck Institute for Meteorology, Hamburg, Germany

Abstract Seasonal prediction is one important element in a seamless prediction chain between weather forecasts and climate projections. After several years of development in collaboration with Universität Hamburg and Max Planck Institute for Meteorology, the Deutscher Wetterdienst performs operational seasonal forecasts since 2016 with the German Climate Forecast System, now in Version 2 (GCFS2.0). Here, the configuration of the previous system GCFS1.0 and the current GCFS2.0 are described and the performance of the two systems is compared over the common hindcast period of 1990–2014. In GCFS2.0, the forecast skill is improved compared to GCFS1.0 during boreal winter, especially for the Northern Hemisphere where the Pearson correlation has increased for the North Atlantic Oscillation index. Overall, a similar performance of GCFS2.0 in comparison to GCFS1.0 is assessed during the boreal summer. Future developments for climate forecasts need a stronger focus on the performance of interannual variability in a model system.

Plain Language Summary Information about the expected departure from the "normal" climatic conditions of an upcoming season would be tremendously valuable for many sectors of society. In Germany, three institutes join their expertise to build a climate forecast system using the Earth system model of the Max Planck Institute for Meteorology. This model describes the atmosphere, land and rivers as well as the ocean and sea ice. The model describes their interactions and is well designed for climate studies on a much longer timescale than a season. Max Planck Institute for Meteorology, Universität Hamburg and the German Meteorological Service Deutscher Wetterdienst have developed the methods those are necessary for such a forecast system and operationally perform the seasonal predictions. This paper compares two versions of our forecast system. The forecast quality during different seasons is particularly investigated. The expectation that the second model system is much better than the first system is not entirely fulfilled. We discuss possible reasons and suggest a stronger focus on the model quality for interannual variability for future model development.

1. Introduction

Over the last two decades, seasonal climate predictions have evolved from a scientific research topic into full-fledged operational systems. Today, seasonal prediction systems are most operationally run at large weather centers like ECMWF (European Center for Medium-Range Weather Forecasts), see Johnson et al. (2019) or the British Met Office (MacLachlan et al. 2015), issuing real-time seasonal forecasts. Thirteen global producing centers (GPC's) currently submit their long-range forecasts to the World Meteorological Organization WMO lead center in South Korea (https://www.wmolc.org/). Since 2011, the German Meteorological Service Deutscher Wetterdienst (DWD), Universität Hamburg (UHH) and the Max Planck Institute for Meteorology (MPI-M) have been developing a system for seasonal climate prediction (Baehr et al., 2015). Since October 2016, the German Climate Forecast System (GCFS) operationally produces seasonal predictions, which are published every month at DWD's homepage. In 2017, DWD became the 13th GPC of the WMO's multimodel ensemble for long-range seasonal forecasts.

At the Max Planck Institute, the Earth-System Model (MPI-ESM; Giorgetta et al., 2013; Mauritsen et al., 2018) is developed. MPI-ESM is tuned over many model-years under pre-industrial conditions (referring to climate conditions of 1850), until it is decided that the best match of the known state of the Earth's climate system is found while keeping the balance of the atmosphere's radiation (Mauritsen et al., 2012).

This balanced model state serves as a basis for any study on introduced anthropogenic changes within the following centuries. With increasing model resolution, a challengingly long computational time is required to achieve an equilibrium model state.

At Universität Hamburg, assimilation and ensemble generation methods are developed for the use in GCFS. DWD adapts and maintains the whole system for operational performance and issues the seasonal forecasts.

In the following, we will describe the configuration of the two versions GCFS1.0 and GCFS2.0, analyze both systems concerning the representation and prediction skill of the North Atlantic Oscillation (NAO), the El Niño Southern Oscillation (ENSO), surface temperature and geopotential height at 500 hPa and compare them with the focus on differences in the model climate and the respective hindcast skill. We will discuss GCFS′ strengths and weaknesses and possible approaches for the future development and conclude with our main findings.

2. Configuration of the Operational Systems

2.1. The Seasonal Forecast System

Performing forecasts with a climate model requires the development of an appropriate workflow. This includes the provision with initial conditions from reference data, the selection of suited methods of assimilation, as well as the generation of an ensemble and the production of hindcasts and forecasts. As an example, Figure 1 sketches the workflow of the seasonal forecast system GCFS2.0 from the generation of the initial conditions to the hindcast and forecast ensembles. This figure and a simple description of the system can also be found under https://www.dwd.de/EN/ourservices/seasonals_forecasts/project_description.html. These tasks will be explained in the following subsections. Details about the two Earth System Model configurations are provided and discussed afterwards, summarized in Table 1.

2.2. Assimilation

Continuous nudging is used to bring the model′s state close to the (re)analyzed state of the climate system. This means a separate simulation runs continuously over the years and months under nudging conditions. Its restart files produced at the end of each month serve as the initial conditions for the retrospective (so-called hindcasts) and current forecasts. The model′s atmosphere is nudged towards vorticity, divergence, temperature, and mean sea level pressure of the (re)analysis data. The model′s ocean is nudged towards temperature, salinity, and sea ice. The methods are the same as described in Baehr et al. (2015). Assimilation for both versions starts in 1979. Data requirements for initial conditions differ for hindcast and real-time forecast.

For the hindcast production both GCFS1.0 and GCFS2.0 use ERA-Interim (Dee et al., 2011) as the atmospheric initial conditions. For real-time forecasts, initial conditions are taken from the analyses of the ECMWF weather forecast model IFS.

Ocean initial states for GCFS1.0 are provided by 3D ocean temperature and salinity of the ECMWF ocean reanalysis ORAS4 (Mogensen et al., 2012) and sea ice concentration from the National Snow and Ice Data Center NSIDC (Fetterer et al., 2002). As reanalysis data are usually not available close to the forecast start date, a special product, the near real-time analysis system ORAS4, provided data for GCFS1.0 forecasts.

Pragmatical considerations for sea-ice data in terms of availability and consistency led to the choice of ORAS5 data (Zou et al., 2017) for all oceanic variables in GCFS2.0. ORAS5 now also provides the near real-time data for GCFS2.0 forecasts.

To start the assimilation run for GCFS2.0 in its very beginning at 1979, we use the decadal assimilation with MPI-ESM-HR from the German decadal climate prediction project MiKlip (Pohlmann et al., 2019; Polkova et al., 2019). The decadal MPI-ESM-HR assimilation begins in 1960, accordingly GCFS2.0 starts in 1979 with an ocean and a land surface already adapted to permanently nudged atmospheric and oceanic conditions.
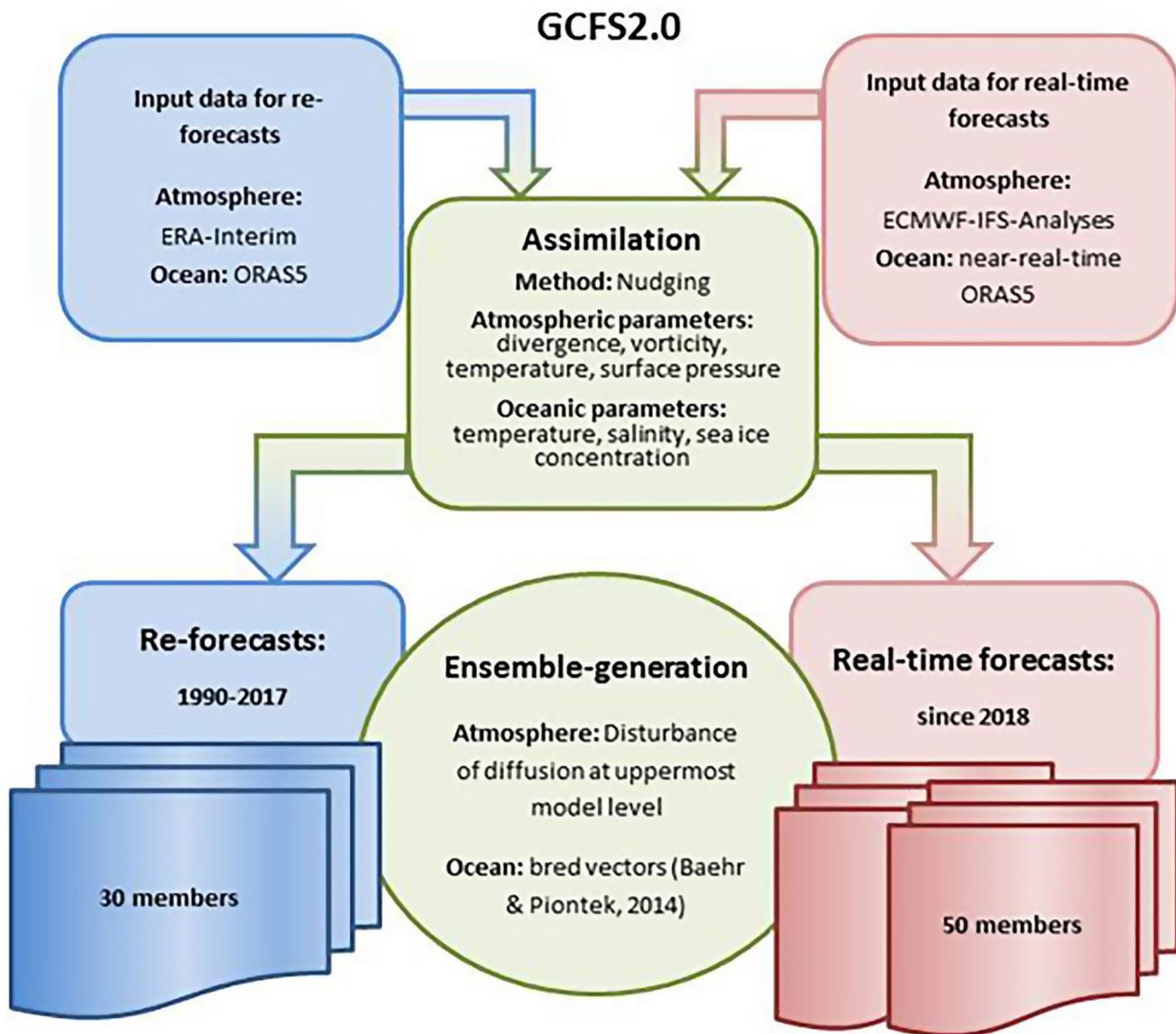
Figure 1. Workflow of GCFS2.0 seasonal forecasts.

2.3. Ensemble Generation

Ensemble members are generated by applying perturbations both to atmosphere and ocean. In the ocean, the ensemble is initialized through bred vectors in all vertical levels (Baehr & Piontek, 2014). The initial ensemble is generated by applying the lagged-day initialization. This means that every member starts on a different day of the preceding month. For instance, on May 1, 1990, the 15 GCFS1.0 members start with conditions taken from days between 1 and 30 April. At the end of this first GCFS1.0 hindcast (April 30, 1991), the ocean restart files of these ensemble members serve as perturbations on the assimilation for the next May hindcast. After approximately two years the bred vectors have lost their memory of the initial lagged perturbations (Baehr et al., 2015). With the increase of GCFS2.0's ensemble size to 30, the duration of every individual hindcast and forecast (and therefore the breeding time) is reduced to 6 months.

Within the atmosphere we apply a simple perturbation of a physical parameter: the perturbed horizontal diffusion coefficient is imposed on vorticity, divergency and temperature at the uppermost level at 0.01 hPa. It forces slightly different atmospheric conditions of the applied member. This small and simple change has a long known impact on the atmosphere model dynamics and was therefore the first choice for an ensemble

Table 1
Overview of GCFS1.0 and GCFS2.0 Configuration

Component	Subcomponent	GCFS1.0	GCFS2.0
Model	Atmosphere/ECHAM	ECHAM 6.1.06p4: T63L47	ECHAM 6.3.04p1; T127L95
	Land surface vegetation/JSBACH	Bulk soil moisture	Five layer soil moisture
	Ocean/MPIOM	MPIOM 1.6.1: GR15	MPIOM 1.6.3; TP04
	Coupler/OASIS3 MCT	Once a day	Hourly
External forcing data	Historical	CMIP5 1981–2005	CMIP6 1981–2014
	Scenarios	CMIP5 RCP 4.5 from 2006 onwards	CMIP6 constant 2014 values from 2015 onwards
Ensemble generation	Atmosphere	Perturbation of uppermost atmospheric layer	Perturbation of uppermost atmospheric layer
	Ocean	Bred vectors over 12 months	Bred vectors over 6 months
Assimilation	Atmosphere	Nudging of ERA-Interim variable fields	Nudging of ERA-Interim variable fields
	Ocean	Nudging ORAS4 and NSIDC variable fields	Nudging of ORAS5 variable fields
Hindcast	Period	1981–2014	1990–2017
	Forecast duration	12 months	6 months
	Ensemble member	15	30
Forecast	Forecast duration	12 months	6 months
	Ensemble member	30	50

generation. The perturbation value varies every fifth member in the hindcast ensemble and every tenth member in the forecast ensemble of GCFS2.0.

2.4. Hindcasts

Hindcasts provide necessary climate statistics for the climate forecasts and the assessment of the model's predictive skill in both deterministic and probabilistic scores. This means that for each of the 12 start months of a year, an ensemble prediction is performed for at least 25 - or more - years of the recent past. Hindcast production inherently consumes a large amount of computational resources.

In GCFS1.0, the hindcast ensemble of 15 members covers the period from 1981 to 2015. The hindcast duration for each start date was 12 months.

The size of the GCFS2.0 ensemble has been increased from 15 to 30. The doubling of the ensemble size together with a higher resolution leads to a considerable increase in computational cost. Therefore, the period of hindcasts and the hindcast duration had to be shortened. The 30 members of GCFS2.0 hindcasts start in 1990 and end in 2017, while performing 6 months of retrospective forecasts for each calendar month in all years. In this way, more than 25 years of hindcasts are achieved.

2.5. Forecasts

The production of the forecasts starts at the beginning of every month, with the exact day depending on the availability of analysis data of atmosphere, ocean, and sea-ice, as explained in the subsection "Assimilation."

GCFS1.0 forecasts ran with 30 ensemble members, having all the same start date at the first of each month, with a forecast duration of 1 year.

With GCFS2.0, 50 ensemble members are integrated over half a year, again all members starting on the first day of the month.

2.6. Communication of Seasonal Predictions

Seasonal predictions are provided as anomaly forecasts with respect to a defined model climate. For instance, ensemble mean anomalies are created by subtracting the lead-time dependent ensemble-mean

model-climate of the chosen reference hindcast period. A bias corrected forecast is obtained by adding the observational based climate of the same reference period to each ensemble member anomaly.

A probabilistic outlook checks how many members cluster in a defined event category. For seasonal forecasts, terciles out of the climatological reference period commonly define the cold/normal/warm or dry/normal/wet events.

Seasonal forecasts are usually averaged over 3 months. The first month is discarded in most cases to account for an initialization shock so that the first forecast period starts with the second to the fourth forecast month, and so on. Therefore, a DJF forecast is issued in November, predicting anomalies for the upcoming December, January and February. Likewise, the May forecasts give the outlook for the months June, July and August.

To account for the moderate skill and huge uncertainty in seasonal predictions, information about the forecast quality is provided along with the forecast itself. Uncertainty estimates are provided via probabilistic representations.

2.7. Evaluation Metrics of Seasonal Hindcasts

A simple but instructive first measure is the difference between the reanalysis and each hindcast data set for a given variable. Decreasing differences between the simulations and a reference data set are usually considered as a success in a correct description of climate processes which is expected to result in improved prediction skill.

To further evaluate the predictive skill of each forecast system, deterministic as well as probabilistic measures are used.

As a conventional deterministic score we apply the Pearson correlation or so-called anomaly correlation coefficient (ACC) on the time series of interannual variations of the ensemble mean hindcasts averaged over 3 months. Here, bias corrected hindcast data in cross-validation are used for the ACC. As such, the anomaly for each member is created by subtracting the model climatology where the corresponding member and year is excluded. The significance of the ACC is tested with a t-statistic. To test the significance of differences in the two ACC's we use a method described by (Siegert et al., 2017), with GCFS1.0 (r_{G1EI}) and GCFS2.0 (r_{G2EI}), also taking into account the correlation between the two versions (r_{G2G1}):

$$T_2 = (r_{G2EI} - r_{G1EI})\sqrt{\frac{(n-1)(n + r_{G2G1})}{2\frac{n-1}{n-3}R + \frac{1}{4}(r_{G1EI} + r_{G2EI})^2(1 - r_{G2G1})^3}}, \tag{1}$$

$$R = (1 - r_{G1EI}^2 - r_{G2EI}^2 - r_{G1G2}^2) + 2r_{G1EI}r_{G2EI}r_{G1G2}. \tag{2}$$

Here, significance is tested by a two-sided test with the threshold of 2.074 for $\alpha = 2.5\%$.

As probabilistic measure for this study, we chose the fair ranked probability skill score RPSS (Ferro, 2014). The fair RPSS is calculated over all defined events or categories (usually 3) of the ensemble hindcasts. The RPSS shows the improvement of using probabilistic forecasts versus using a climatological value. The score is called *fair* as it is adapted to the finite ensemble size. The perfect score is 1, values below 0 denote that the climatology performs better than the respective hindcast ensemble. Further metrics to estimate the reliability, resolution, and sharpness of the ensemble system are also calculated for GCFS but not shown here for lack of space. For more details on skill measures and scores we refer to Wilks (1995).

2.8. Model Configuration

The first version of GCFS1.0 was based on MPI-ESM-LR (Giorgetta et al., 2013), with an atmosphere resolution of T63 (corresponding to approximately 150 km at around 50°N), 47 levels reaching up to 0.01 hPa and an ocean resolution of nominally 1.5° in the horizontal with 40 levels in the vertical down to approximately 5,000 m (Jungclaus et al., 2013). The ocean model is connected to a dynamic and thermodynamic

sea-ice model. Coupling between ocean and atmosphere was set to once a day for GCFS1.0. The land and vegetation model JSBACH also hosts a hydrological runoff model. In GCFS1.0, MPI-ESM uses the external forcing like greenhouse gases, ozone, and aerosols based on phase 5 of the Coupled Model Intercomparison Project CMIP5 (Taylor et al., 2012) for historical data and future scenarios. Together with the prescribed solar irradiance these data also account for the solar cycle. Stratospheric aerosol data provide information about volcanoes (Giorgetta et al., 2013). The historical period within CMIP5 ended in 2005. Therefore, all simulations of GCFS1.0 starting from 2006 onwards used the RCP4.5 scenario for the external forcing.

The version GCFS2.0 is based on the MPI-ESM-HR (Mauritsen et al., 2018; Müller et al., 2018) with a T127 spectral resolution in the atmosphere corresponding to approximately 70 km at around 50°N, with 95 levels covering the same vertical column up to 0.01 hPa and an ocean resolution of nominally 0.4° in the horizontal and, similar to GCFS1.0, 40 levels in the vertical. A number of developments in the atmospheric processes like radiation, clouds and convection has been applied while ocean physical processes remained largely unchanged. A complete description of all model developments of the new and higher resolved Earth-System Model is provided by Mauritsen et al. (2018). We highlight here the increase of the coupling frequency between ocean and atmosphere, which in GCFS2.0 takes places on an hourly basis. Further, the land and vegetation model JSBACH in MPI-ESM-HR includes vertical soil moisture transport (Hagemann & Stacke, 2015). For the low-resolution version MPI-ESM-LR it has been shown, that this contributes to an improvement of European summer temperatures (Bunzel et al., 2017). For GCFS2.0, we use external forcing from phase 6 of the Coupled Model Intercomparison Project CMIP6 (Eyring et al., 2016), where the historical period has been extended until 2014. Scenario data were not ready for use for MPI-ESM-HR before mid-2018, therefore the external forcing has been set constant starting in 2015 up to present time. Table 1 gives an overview of both systems.

To give an idea of the different model behavior purely due to the changed model configuration and physics, Figure 2 shows biases of 2 m temperature of the two climate model versions CMIP5 MPI-ESM-LR and CMIP6 MPI-ESM-HR with respect to ERA-Interim for the time range 1990–2014 obtained from historical experiments. The so-called historical experiments are uninitialized model simulations, starting from 1850 until 2014, where the model climate is controlled by the changing external fields like greenhouse gases, aerosol and ozone. Displayed are ensemble means from the 10 member ensemble for December/January/February DJF in Figure 2 (top) and June/July/August JJA Figure 2 (bottom).

Over land, the warm bias of the Amazon basin is considerably reduced in the CMIP6 model version, which we attribute to the improved soil moisture behavior. Seasonal differences also show up, for instance the reduced cold boreal summer bias in northern Russia and northern Africa in the new and higher resolved model system. The bias pattern over northern America also changes between the model versions and the two different seasons, but no clear error reductions are visible there. Europe exhibits a cold bias in the CMIP6 version for DJF and JJA, while the CMIP5 version produced a warm DJF bias. The European cold JJA bias is slightly reduced in the CMIP6 version. Another cold bias region is visible in the North Atlantic. The pattern remains similar between the two versions and appears to be stronger in winter. The two ocean grids GR15 and TP04 (see Table 1) have a very similar resolution setting over the North Atlantic. Therefore, there is no visible effect of an improvement due to a better resolved ocean at the ocean surface. As shown in Gutjahr et al. (2019), the MPI ocean model can represent the pathway of the North Atlantic Current much better at even higher horizontal eddy resolving resolution including improved physics with regard to vertical mixing.

In the Southern Hemisphere, a warm bias is visible around the coast of Antarctica for both model versions and both seasons. The bias is slightly reduced in the CMIP6 version, as cloud errors have been alleviated (compare with Figure B6 in Mauritsen et al. (2018)) and possibly as well due to the higher resolution which was also shown in Müller et al. (2018), there Figure 2. Still, sea-ice around Antarctica is not well represented, particularly too little ice in summer (not shown).

Although some of these biases can be addressed during data assimilation, most will re-emerge in the seasonal hindcasts and forecasts.

Another striking features of the CMIP6 version for both seasons are the much warmer tropical oceans compared to the reanalysis and the reduced cold tongue in the tropical Pacific in Figures 2b and 2d. Although
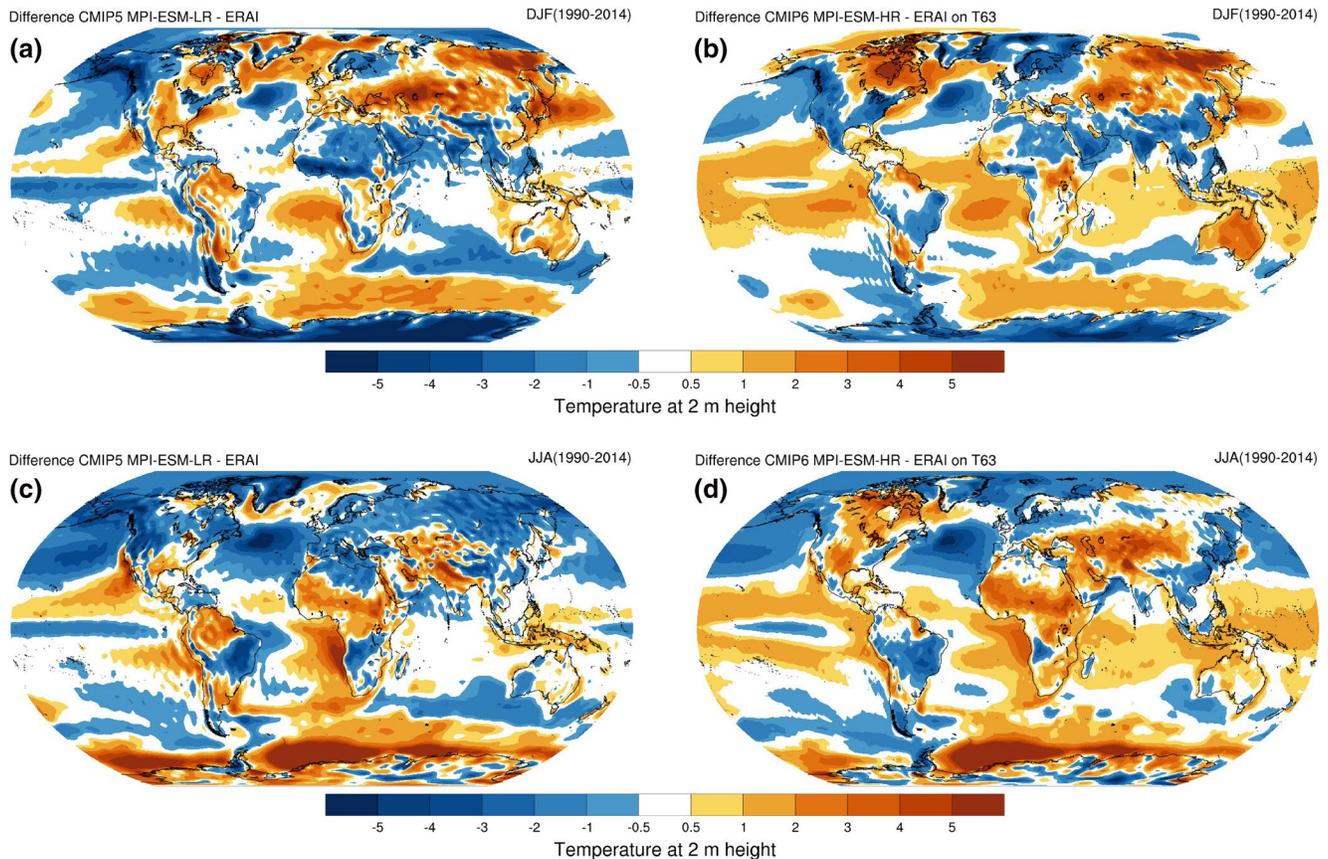
Figure 2. Difference of the 2 m temperature between the historical experiment of CMIP5 MPI-ESM-LR (left) and CMIP6 MPI-ESM-HR (right) to ERA-Interim for DJF (top) and JJA (bottom) between 1990 and 2014.

this error pattern has not disappeared completely, it is much smaller for the DJF months than for the JJA months.

An issue with the ocean mixing has been reported in Mauritsen et al. (2018). A "bugfix" was implemented and tuned already for the low resolution version CMIP6 MPI-ESM-LR, but for the high resolution version MPI-ESM-HR it was decided to leave out this further tuning connected with the ocean correction. To investigate the impact of the ocean code correction onto the new MPI-ESM-HR climate for seasonal forecasts an experiment was set up. Here, MPI-ESM-HR is run for approximately 100 years under pre-industrial control conditions with correct ocean mixing in order to allow a spin-up for the ocean. Afterwards, a historical experiment simulates the time range from 1850 up to 2014. However, the comparison between the original historical run and the bug-fixed version did not show the desired result of a cooler tropical belt. The revised ocean mixing in MPI-ESM-HR revealed a small improvement (not shown here) but no change of error pattern in the tropical Pacific similar to the change seen in Figure 2. Except for Arctic regions, temperature differences between these two MPI-ESM-HR versions remain mostly below 1 K and are hard to recognize. The reason for the tropical warm bias is thus suspected to derive from the revised atmospheric parameterizations in radiation and cloud cover. Required diagnostics and tuning experiments are currently beyond available resources.
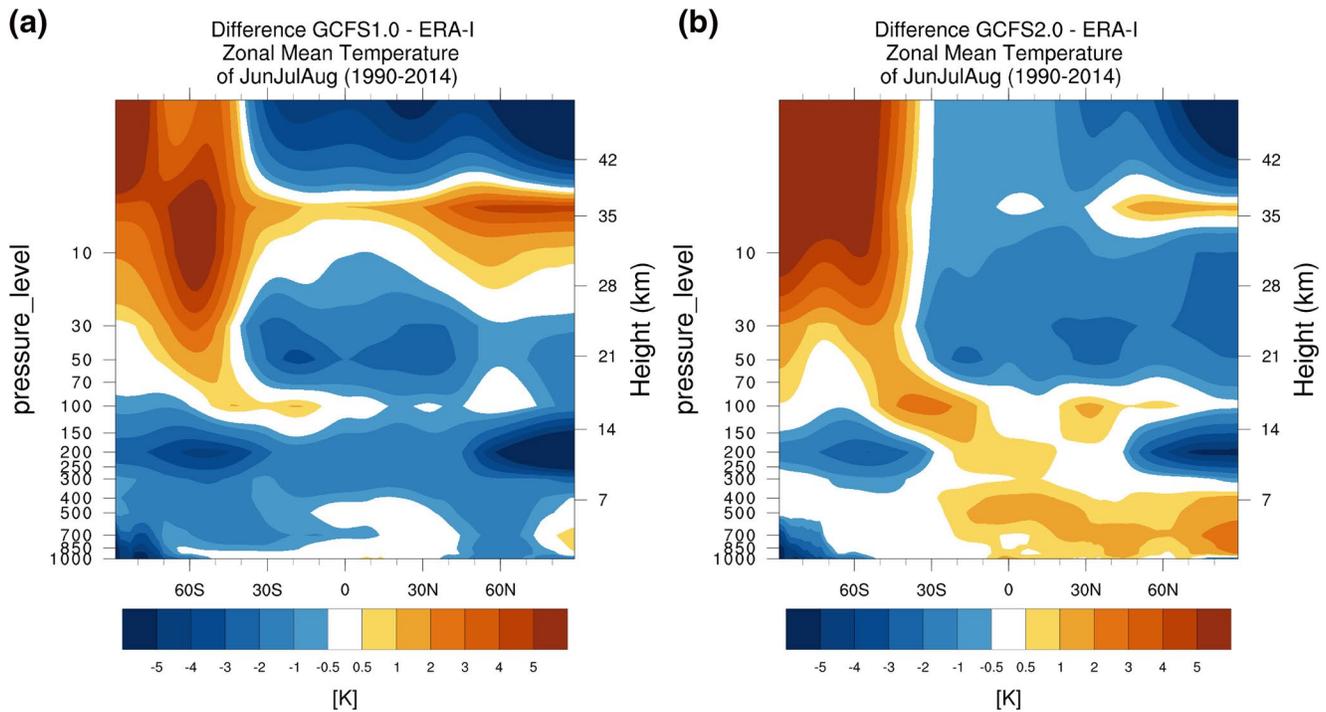
Figure 3. Temperature differences of (a) GCFS1.0 ensemble mean and (b) GCFS2.0 ensemble mean to ERA-Interim of JJA hindcasts during 1990–2014.

3. Comparison of GCFS1.0 and GCFS2.0 Hindcasts

3.1. Comparison Setup

With May and November we use two of the 12 start months to investigate the hindcast performance of the two forecast systems. To have a common hindcast period of GCFS1.0 and GCFS2.0 the reference time of 1990–2014 is chosen. We focus our assessment on temperature and geopotential fields. Mean sea level pressure is used for the evaluation of the NAO prediction. The reference dataset is the ERA-Interim reanalysis except for the evaluation of the Niño3.4 regions, where the NOAA Optimum Interpolation Sea Surface Temperature Version 2 NCEP OIv2 is used.

3.2. JJA Mean State and Hindcast Skill

We present the mean state of the models by looking at the vertical structure of the atmosphere as well as at the surface and the level of 500 hPa. We consider the ensemble mean and time mean of the respective 3 month hindcasts over the time period of 1990–2014.

Figure 3 visualizes in a latitude-height-plot the bias of temperature of GCFS1.0 (left) and GCFS2.0 (right) with respect to ERA-Interim. The vertical and the latitudinal temperature structure in the two panels is quite different. In stratospheric levels around/above 35 km a dipole feature of the boreal summer stratospheric temperature bias of GCFS1.0 (Figure 3a) is visible. This feature is weakened in GCFS2.0 (Figure 3b), where a higher vertical resolution exists, a new ozone data base for CMIP6 is used and the non-orographic gravity waves are newly tuned. However, the middle atmosphere is characterized by a cold bias in the Northern Hemisphere summer and a strong warm bias in the Southern Hemisphere winter. The comparison shows that the tropospheric cold bias in the lower layers in GCFS1.0 during the JJA months is reduced in GCFS2.0, revealing almost a bias-free Southern Hemisphere. Still, the extra-tropical cold biases in 200 hPa indicate a wrong position of the tropopause in both versions. In the GCFS2.0 Northern Hemisphere, a considerable warm bias in the middle troposphere extends from the northern polar latitudes into the tropics. The maximum of the positive bias is placed above the North Pole while in the lowermost layers the cold bias
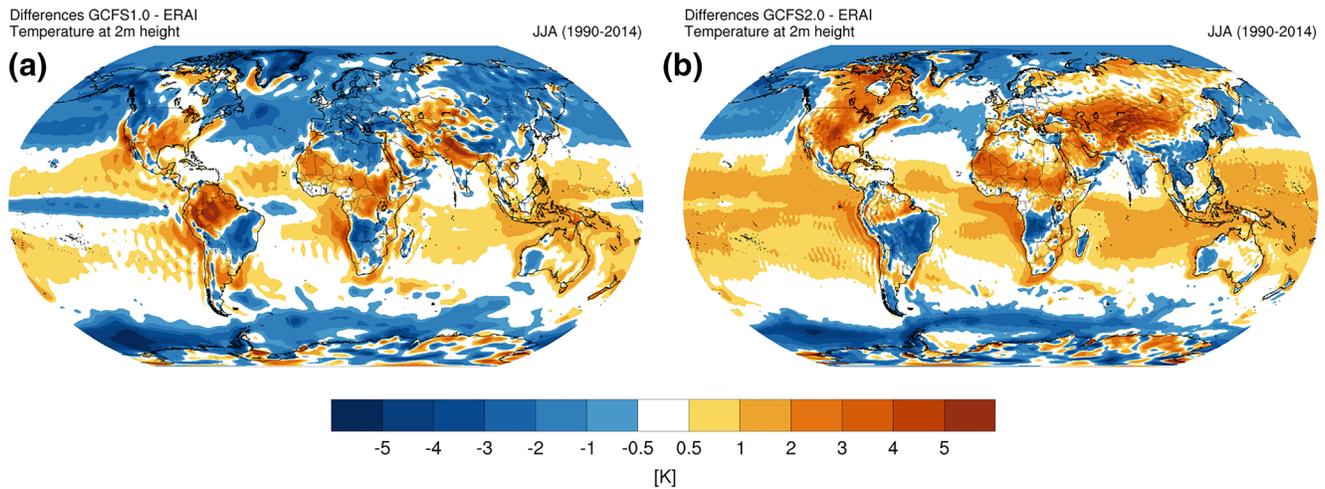
Figure 4. Difference of the 2 m temperature hindcasts (a) GCFS1.0 and (b) GCFS2.0 to ERA-Interim for JJA hindcasts in 1990–2014.

of the Arctic remains. The warm bias Northern Hemisphere region is accompanied by a positive humidity bias (not shown here).

JJA hindcast temperature biases at 2 m height are displayed in Figure 4. The cold bias of Northern Hemisphere land masses in GCSF1.0, which can also be seen in the previously mentioned 10 member ensemble of Baehr et al. (2015), is greatly reduced. However, in some regions it is replaced by a warm bias in GCFS2.0, for example, over North America. The error in the Amazon basin is strongly reduced in GCFS2.0, which we attribute to a feature of the new soil moisture parameterization. The cold tongue in the tropical Pacific, present in GCFS1.0, completely vanished in GCFS2.0. This seems to be due to a strong impact of ORAS5-nudging. Additionally, the tropical Pacific now shows a strong warm bias in the upwelling region west of the South American and South African coasts. The calculation of the Gaussian weighted, globally averaged RMSE, separately over land and ocean, for both forecast versions is shown in Table 2. This table confirms that a redistribution of the error patterns results in similar RMSE values in GCFS2.0 during boreal summer.

We investigate the two systems with respect to their forecast skill in terms of ACC and fair RPSS in Figure 5. Figure 5a shows, that for central Europe no hindcast skill in GCFS2.0 temperature can be expected, while the situation is better in North America, the Mediterranean and some Asian regions, where significant correlation values appear, represented by dots. In the middle troposphere, the ACC for the GCFS2.0 geopotential height (Figure 5b) shows negative values over Europe. Again, western US, Greenland and Central Asia exhibit significant positive skill outside the tropical regions.

The middle panels, Figures 5c and 5d, show the change in skill from GCFS1.0 to GCFS2.0. For the inter-model comparison GCFS2.0 is mapped onto the coarse grid of GCFS1.0. If regions appear in reddish colors GCFS2.0 is superior over its predecessor GCFS1.0. With the applied method from Siegert et al. (2017) significant patterns also appear where the difference in the two correlations ($r_{G2EI} - r_{G1EI}$) is very weak. These places represent regions, where the correlation between GCFS1.0 and GCSF2.0 r_{G1G2} is very high so that even small differences matter.

Table 2
Gaussian Weighted Globally Averaged RMSE of 2 m Temperature for JJA

Forecast system	RMSE [K] land	RMSE [K] ocean
GCFS1.0	0.95	0.55
GCFS2.0	0.97	0.55

At the surface relevant improvements for the temperature correlation can be seen over Alaska, the west Siberian Plain, the Amazon and the western tropical Pacific region. A slight improvement is gained in northern Europe. At 500 hPa the correlation for the geopotential height is now stronger over Greenland and parts of Siberia as well as over parts of Antarctica. The negative skill over Europe is a pattern which has not changed much in comparison to GCFS1.0, but some significant improvements can be stated over Scandinavia.
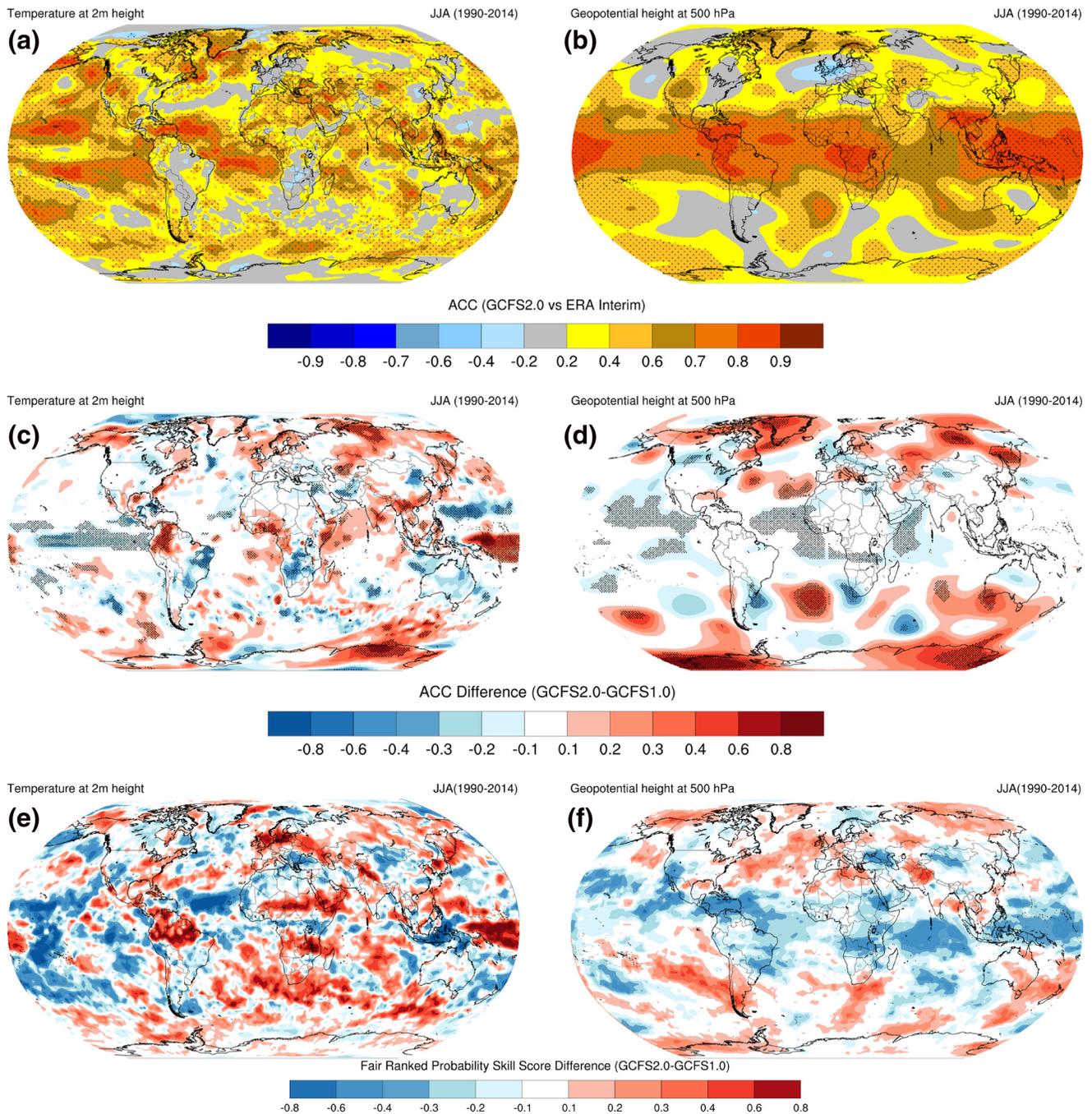
Figure 5. Top panel: Anomaly correlation coefficient (ACC) of hindcasts for June, July, August with respect to ERA-Interim during 1990–2014: (a) 2 m temperature and (b) geopotential height at 500 hPa pressure level. Dots represent significant values at $\alpha = 5\%$. Middle panel: the difference of ACC skill between GCFS1.0 and GCFS2.0 for (c) 2 m temperature and (d) geopotential height at 500 hPa. Dots represent significant values at $\alpha = 2.5\%$ derived from Equation 2. Bottom panel: difference of GCFS1.0 and GCFS2.0 for the Fair Ranked Probability Skill Score RPSS for (e) 2 m temperature and (f) geopotential height at 500 hPa.

The difference in the probabilistic measure RPSS between the two model systems is shown in the bottom panels of Figure 5, again for temperature on the left in Figure 5e and geopotential height on the right in Figure 5f. As before, the red-colored regions highlight the domains where the probabilistic hindcasts of GCFS2.0 for all event categories are better than the GCFS1.0 ensemble. The differences between the systems are stronger at the surface than in 500 hPa. GCFS2.0 probabilistic hindcasts have improved over the North and Baltic sea
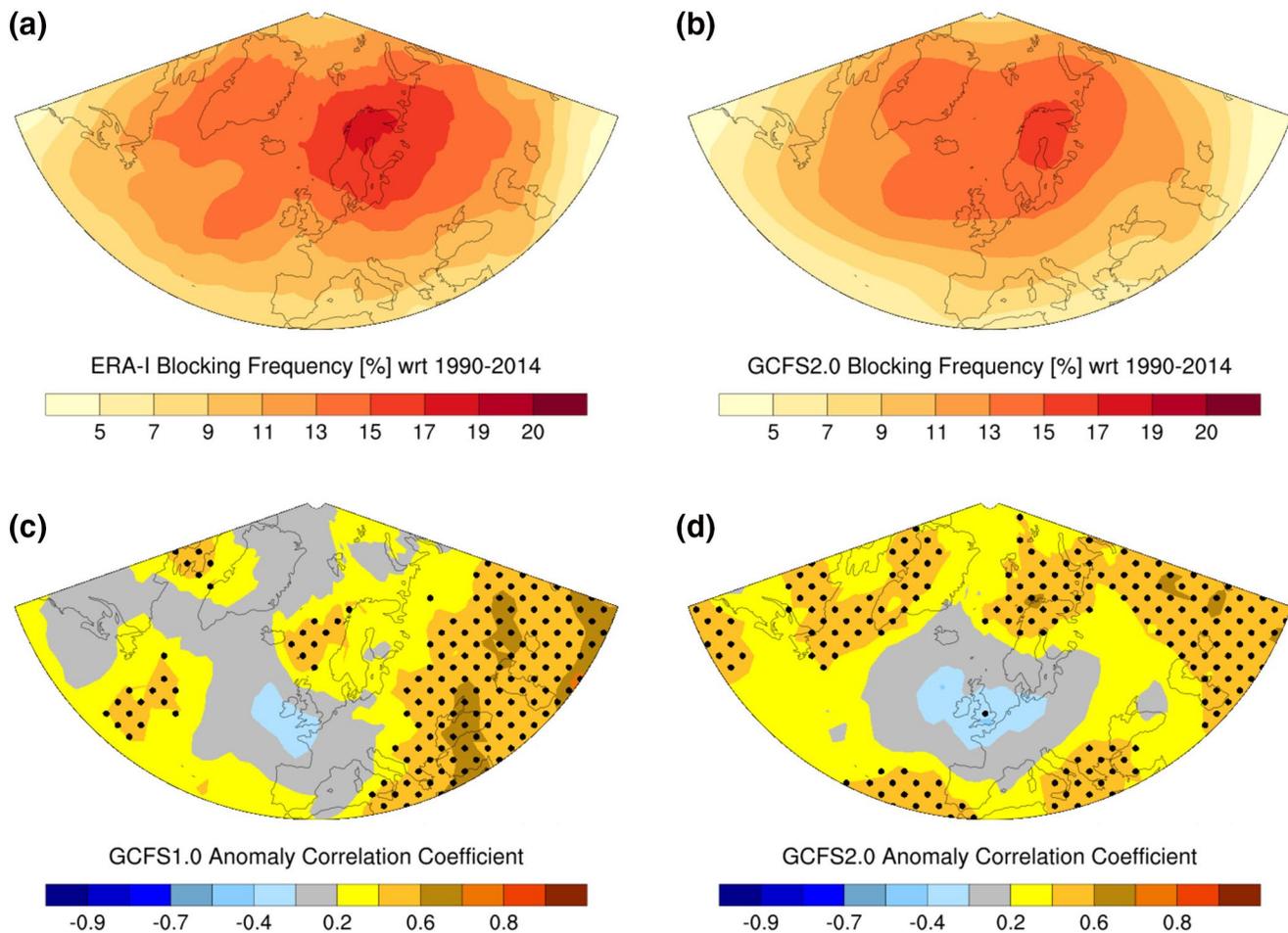
Figure 6. Analysis of blocking frequency in JJA during 1990–2014 for (a) ERA-Interim reanalysis, (b) GCFS2.0 ensemble mean, (c) ACC for GCFS1.0 vs. ERA-Interim, and (d) ACC for GCFS2.0 vs. ERA-Interim. Dots represent significant values at the 5% significance level.

and its surroundings, in the Sahel zone and again over the Amazon region. Degradation is seen for the central tropical Pacific, tropical Atlantic and the Indonesian Archipelago. For geopotential height at 500 hPa probabilistic hindcasts are now slightly worse around the tropics, while the North Atlantic shows neutral to slightly improved behavior. The structure and hindcast skill of one of the prominent European summer features, blocking, is shown in Figure 6. Blocking is diagnosed from daily values of geopotential height at 500 hPa of the hindcasts started in May by using a combination of two methods based on Tibaldi and Molteni (1990) and Barriopedro et al. (2010). The ERA-Interim reanalysis (Figure 6a) shows that the European blocking has its maximum over northern Scandinavia. The region where events are present for more than 15% of the summer days during the considered time range extends from Greenland to the Ural Mountains and from Spitsbergen down south to Poland and Ukraine. Please note, we compare the seasonal frequency of summer blocking events but not their correct timing during the summer. While the general shape and the location of the central core of European summer blockings match quite well between reanalysis and forecast systems, the extension and especially the amplitude of GCFS2.0 blockings (Figure 6 b) are smaller. Panels c) and d) of Figure 6 show the ACC of the two versions with dots representing significant values at the 5% significance level. In general, GCFS2.0 better describes the northern and western blocking events, while GCFS1.0 performed better in south-eastern Europe. Over central Europe, the situation is not improved by GCFS2.0. The underestimation of this phenomenon is, however, not unexpected, as blocking processes are known to evolve properly only at horizontal grid resolutions of about 40 km as shown in, for example, Jung et al. (2012). A further reason of the underestimation of blocking could arise from the cold bias in the North-Atlantic, as has been described for winter blockings by Scaife et al. (2011). This cold bias is considerably reduced in GCFS2.0 but still present. The
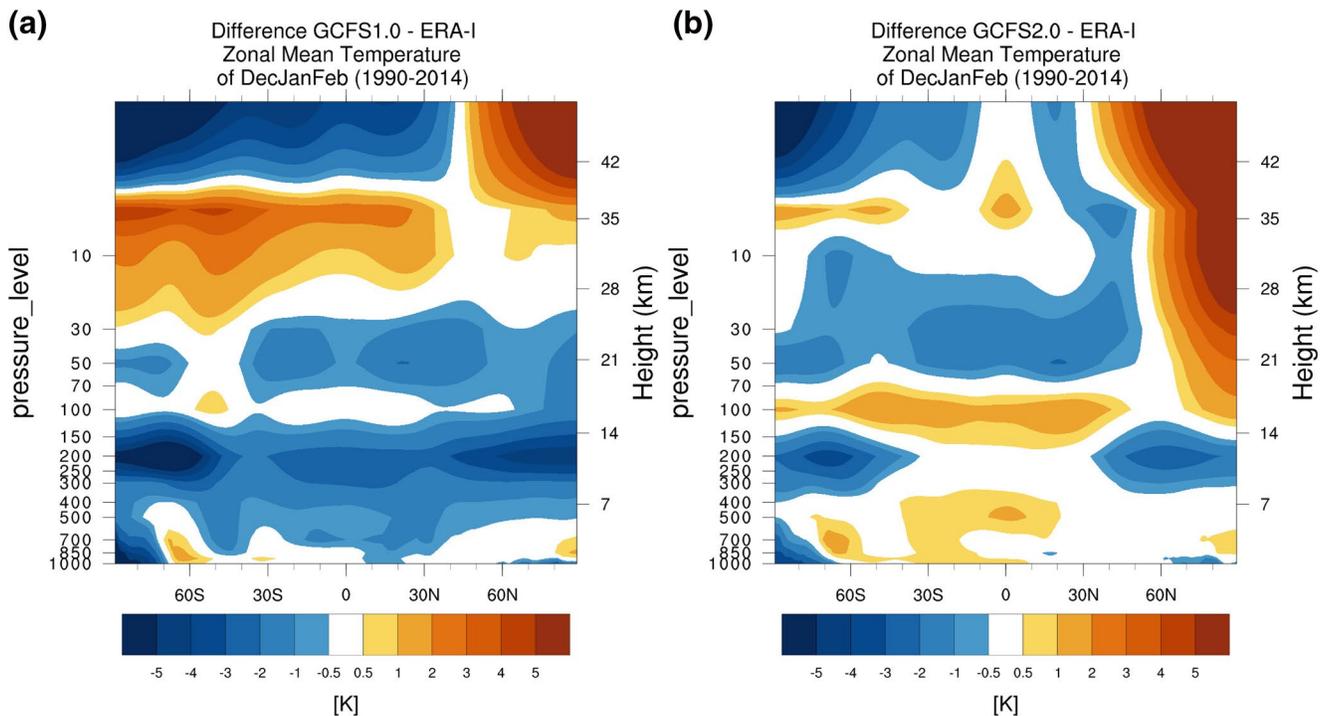
(a)

(b)

Figure 7. Temperature differences of (a) GCFS1.0 ensemble mean and (b) GCFS2.0 ensemble mean to ERA-Interim with respect to DJF hindcasts during 1990–2014.

ocean resolution of 0.4° of GCFS2.0 is still not high enough to resolve the processes the North Atlantic current and to place it at the correct position in the North Atlantic.

3.3. DJF Mean State and Hindcast Skill

As before, the ensemble and time means of GCFS1.0 and GCFS2.0 are compared by using the corresponding ERA-Interim period of December, January and February (DJF) 1990–2014, computed from the November start date. The year is related to the start of the season, namely December.

Again, we show first the vertical structure of the temperature differences in Figure 7 for the DJF period. For GCFS1.0 (Figure 7a), a strong cold bias is seen in the upper troposphere extending with weaker values up into the stratosphere. In the GCFS2.0 troposphere (Figure 7b), the cold bias is considerably reduced. In tropical and subtropical latitudes, a minor warm bias is now present, with the maximum at the tropopause.

The middle atmosphere of the GCFS1.0 Southern Hemisphere is overly warm up until 35 km, while above a strong cold bias is visible. In contrast, the winter stratosphere in the Northern Hemisphere at the same height appears to be too warm. This dipole structure is strongly reduced in GCFS2.0. However, the warm bias in the winter polar stratosphere extends now from about 100 hPa up to 1 hPa. The polar vortex is too strongly decelerated due to resolved and parameterized wave activity.

At the surface, the bias distribution has also changed as is displayed in Figure 8. The tropical oceans now show a general warm bias for GCFS2.0 (8b), as seen before in the historical experiment (Figure 2). In the Pacific, the cold tongue again vanishes as simulations start from an analyzed climate state. Other strong biases from the GCFS1.0 are now also considerably reduced, such as the warm bias over Europe, the Amazon region and southern Africa or the cold bias over northern Africa. For boreal winter also the globally averaged RMSE is reduced in GCFS2.0, as shown in Table 3.

With these improvements in the higher resolved version GCFS2.0 has a better representation of the jet in the storm track region (see Figure 10 from Müller et al. (2018)). This leads to a good skill pattern in ACC of 2 m temperature as seen in Figure 9a and also for the geopotential height at 500 hPa pressure level (Figure 9b).
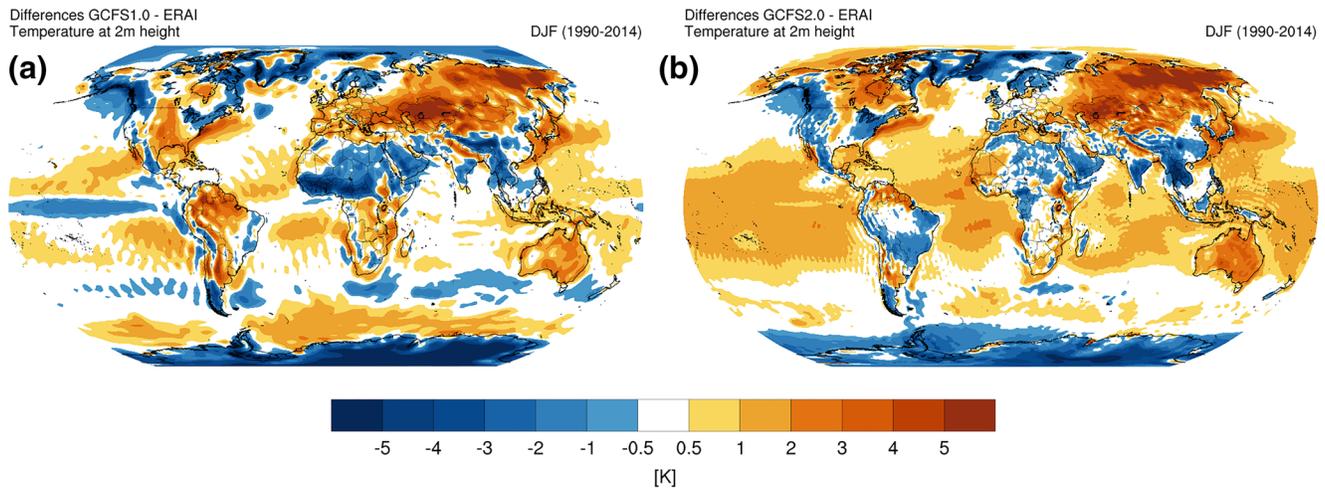
Figure 8. Difference of the T2m hindcasts (a) GCFS1.0 and (b) GCFS2.0 to ERA-Interim for DJF hindcasts between in 1990–2014.

While ACC values over Europe in DJF are not apparent, the skill in the Arctic region, over Greenland and northern America shows significant values up to 0.8 (Figure 9a), whereas in GCFS1.0 the skill did not exceed values around 0.4 (not shown here). However, directly south of Greenland appears an oceanic region with negative skill. This is a new pattern in GCFS2.0 and is most probably not a model feature but could stem from assimilating the ocean reanalysis data ORAS5 from ECMWF. The same negative Pearson correlation appears in the DJF hindcast skill of ECMWF's forecast model System 5 (Johnson et al. , 2019), see their Figure 19a. The authors also discuss ORAS5 as the reason of skill degradation. At the level of 500 hPa again the strongest positive values are found in the tropics. Apart from this, regions with considerable positive and significant skill are found over the eastern North Pacific extending into North America, as well as over Greenland and Arctic regions. Also, the North Atlantic shows a patch of significant positive skill.

The middle panels of Figure 9 highlight as before the change in skill in the ACC between the two versions GCFS1.0 and GCFS2.0, with dots over the significant regions. Again, significant regions also appear, where the difference in the two correlations ($r_{G2EI} - r_{G1EI}$) is very weak but the correlation between GCFS1.0 and GCSF2.0 r_{G1G2} is very high. Considerable skill has been gained at the surface, represented by the 2 m temperature (Figure 9c). Greenland and large parts of Eurasia benefit from the new version. Europe partly gains skill, especially for the very north and the southern regions and partly loses skill, especially over central Europe. A substantial gain in skill is further evident for the geopotential height in the storm track level in the North Atlantic, Arctic regions (Figure 9d), as well as over Eurasia. The significant improvement over central Europe is gained by replacing negative ACC values with weak positive ones.

The bottom panels in Figure 9 present the difference between the two systems in terms of the probabilistic hindcast skill score RPSS as previously shown in Figure 5. Skill differences are stronger at the surface for temperature than for the middle troposphere in 500 hPa. During DJF GCFS2.0 probabilistic hindcasts are more skillful for eastern and southern Europe as well as eastern Russia. The skill amendment over Europe is seen also at 500 hPa. A region where GCFS2.0 forecasts are worse than its predecessor is the central tropical Pacific. The strongest gain in skill is seen over central Africa extending into the Indian Ocean. Improvements over the ocean are also visible for the subtropical south-eastern Pacific and the Agulhas Basin.

The improved conditions in the Northern Hemisphere show up as well in a skillful prediction of the NAO index in GCFS2.0 for the upcoming boreal winter season (Figure 10). The NAO skill does not directly suffer from the SST problems in ORAS5 reanalysis mentioned above, because it arises from a different SST-region in the North Atlantic, and also depends on northern hemispheric sea ice, snow cover and stratospheric temperatures. The NAO index has been calculated using an empirical orthogonal function (EOF) from mean sea level pressure as in Dobrynin et al. (2018). The

Table 3
Gaussian Weighted Globally Averaged RMSE of 2 m Temperature for DJF

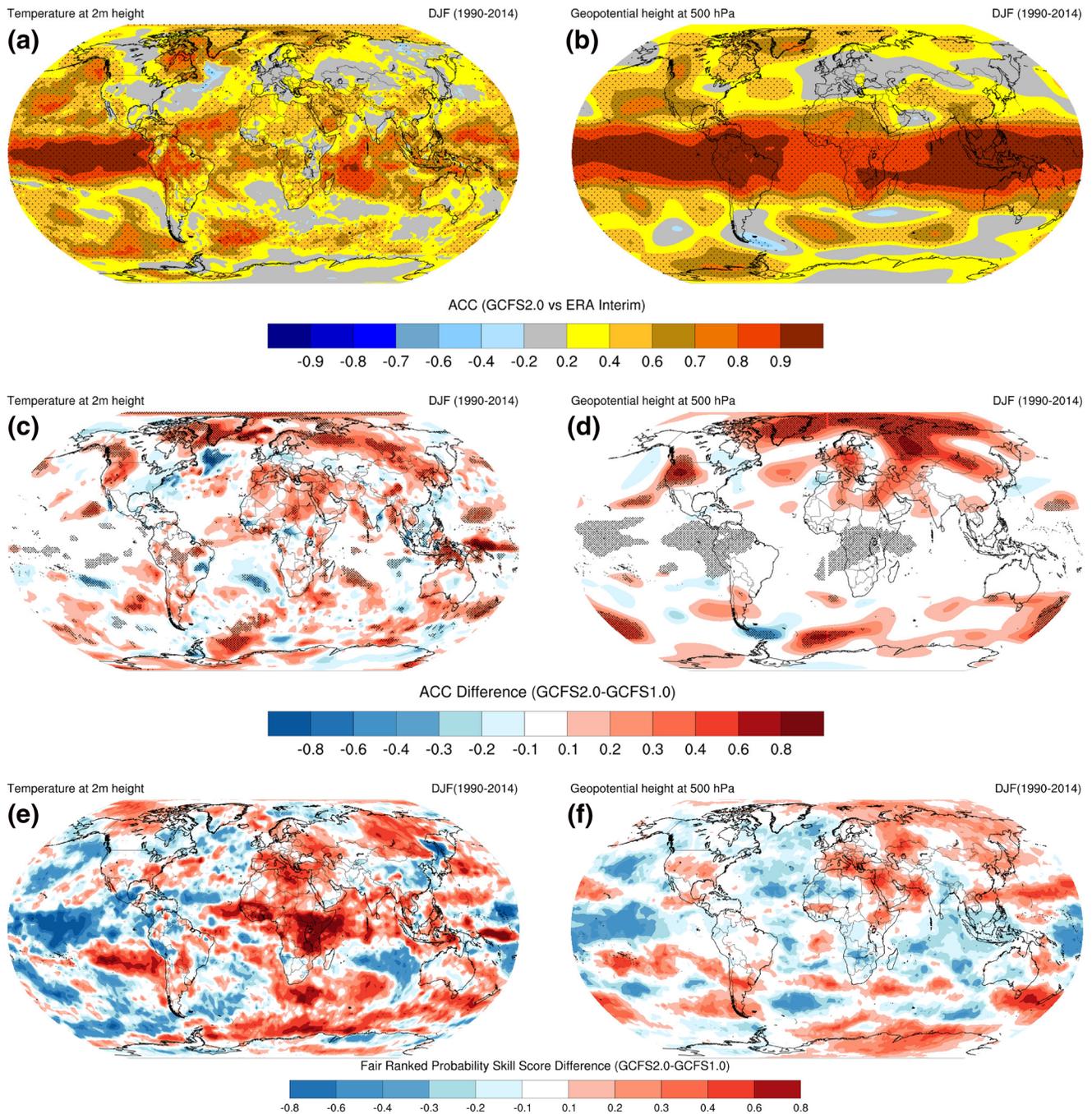| Forecast system	RMSE [K] land	RMSE [K] ocean
GCFS1.0	1.41	0.59
GCFS2.0	1.36	0.56

Figure 9. Top panel: Anomaly correlation coefficient (ACC) of hindcasts for December, January, February with respect to ERA-Interim: (a) 2 m temperature and (b) geopotential height at 500 hPa pressure level. Dots represent significant values at $\alpha = 5\%$. Middle panel: the difference of ACC skill between GCFS1.0 and GCFS2.0 for c) 2m temperature and d) geopotential height at 500 hPa. Significance is represented by dots at $\alpha = 5\%$ derived from Equation 2. Bottom panel: the difference of GCFS1.0 and GCFS2.0 for the Fair Ranked Probability Skill Score for (e) 2 m temperature and (f) geopotential height at 500 hPa.

monthly values have been normalized by the standard deviation of the monthly NAO index time series from 1950 to 2000. Comparing to the previous version GCFS1.0 with a not significant NAO prediction skill of 0.21, the GCFS2.0 increased the NAO skill up to 0.40, which is statistically significant at the 5% significance level, although the difference in NAO skill between the two systems is not significant. Further, the NAO skill for a reduced 15-member GCFS2.0 ensemble is still considerably higher than in GCFS1.0. The correlation means over all combinations for 15 out of 30 members in GCFS2.0 is 0.34, with a range between 0.01 and 0.62 dependent
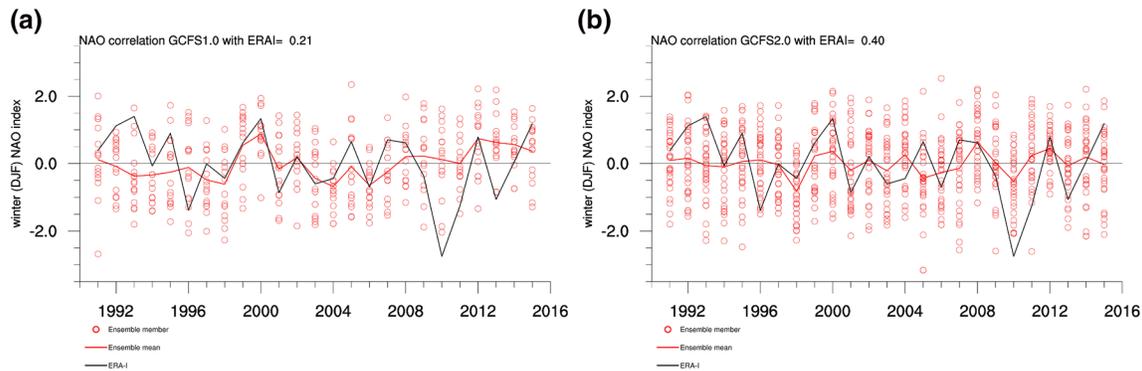
Figure 10. NAO time series of ERA-Interim, model ensemble mean and the single ensemble members averaged over December, January, February during 1990–2014, (a) GCFS1.0 and (b) GCFS2.0. Time series are normalized by the standard deviation. The labeled years refer to January of the considered DJF.

on the selected members. This highlights that improvements of the model dynamics and physics have large potential for better long-range forecasts in the mid-latitudes as shown by Scaife et al. (2014).

3.4. ENSO Hindcasts

The assessment of ENSO hindcasts in Figure 11 for the Niño3.4 region shows a strong seasonal dependence of skill in GCFS2.0. However, when considering all 12 start months predictive skill is reduced in GCFS2.0. The general structure of the difficult hindcast start months March, April, May and June can also be seen in GCFS1.0 and is also known in other models as "spring predictability barrier" (see e.g., Wang-Chun Lai et al. (2017)), making ENSO predictions difficult for forecast systems in general. This feature has not been improved in GCFS2.0. However, GCFS2.0 performs with a comparable forecast quality to GCFS1.0 in all other months from July to December.

Johnson et al. (2019) also report a warm bias of the ECMWF's seasonal forecast model System 5 during JJA hindcasts in the eastern Pacific basin (their Figure 1d). The warm structure of System 5 is stronger pronounced north of the equator and much more confined to the South American coast. Their ENSO predictive quality improved with the higher resolution of the model.

4. Conclusions

To summarize, we have shown that the second version of the GCFS has some improvements over its predecessor. However, a version change does not necessarily lead to improvement everywhere and for every variable, which is certainly true here. An overall similar behavior of GCFS2.0 in comparison to GCFS1.0 has been assessed during JJA. Skill degradation of GCFS2.0 in JJA forecasts is prominent in the tropical Pacific and in the skill for ENSO. During DJF, hindcast quality is improved in GCFS2.0 due to the increased resolution and revised physical parameterizations, especially for Northern Hemisphere, where, for example, NAO skill for the winter months increased.

With this version, we have again learned that increasing the model resolution is not per se a solution to many forecast problems as it requires a lot of intense work and evaluation of the model physics on the new grid. Similar challenges are observed for the new system 5 at ECMWF (Johnson et al., 2019), where many issues arise with the new resolution. Scaife et al. (2019) even recommend to invest more into the ensemble size, vertical resolution or ocean resolution than in increasing the atmospheric horizontal resolution. As can be seen from our results, any changes to the model grid need careful adaptation of the model physics.

For a future system, a comprehensive performance testing is needed depending on different time-scale applications of this model (Schmidt et al., 2017), besides taking into account a well-tuned climate and climate sensitivity of the Earth-System Model, which is the key for climate experiments and climate projections. In this way, processes active in different seasons or timescales can be accounted for during the model tuning.
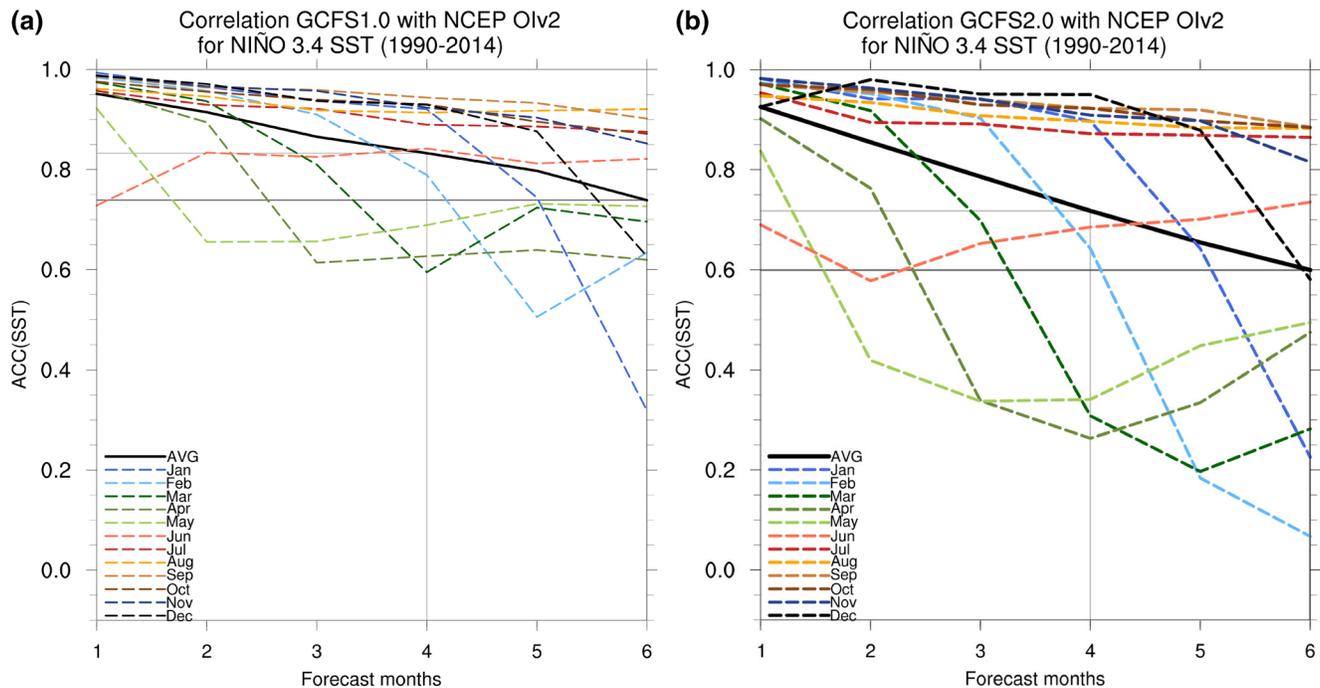
Figure 11. Anomaly correlation of the SST forecasts for the ENSO 3.4 region for (a) GCFS1.0 and (b) GCFS2.0, both calculated with respect to the NCEP reanalysis for all 12 start months and all 6 lead months in 1990–2014.

Data Availability Statement

Hindcast data used for this study and scripts to generate the figures are available by request to the first author under https://doi.org/10.5281/zenodo.3697080. GCFS2.0 data used for this study are available on the C3S climate data store https://climate.copernicus.eu/seasonal-forecasts.

References

Baehr, J., Fröhlich, K., Botzet, M., Domeisen, D. I. V., Kornblueh, L., Notz, D., et al. (2015). The prediction of surface temperature in the new seasonal prediction system based on the mpi-esm coupled climate model. *Climate Dynamics*, *44*(9-10), 2723–2735. https://doi.org/10.1007/s00382-014-2399-7

Baehr, J., & Piontek, R. (2014). Ensemble initialization of the oceanic component of a coupled model through bred vectors at seasonal-to-interannual timescales. *Geoscientific Model Development*, *7*(1), 453–461. https://doi.org/10.5194/gmd-7-453-2014

Barriopedro, D., Garcia-Herrera, R., & Trigo, R. (2010). Application of blocking diagnosis methods to general circulation models. Part i: A novel detection scheme. *Climate Dynamics*, *35*(7-8), 1373–1391. https://doi.org/10.1007/s00382-010-0767-5

Bunzel, F., Müller, W. A., Dobrynin, M., Fröhlich, K., Hagemann, S., Pohlmann, H., et al. (2017). Improved seasonal prediction of european summer temperatures with new five-layer soil-hydrology scheme. *Geophysical Research Letters*, *45*(1), 346–353. https://doi.org/10.1002/2017GL076204

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Dobrynin, M., Domeisen, D. I. V., Müller, W. A., Bell, L., Brune, S., Bunzel, F., et al. (2018). Improved teleconnection-based dynamical seasonal predictions of boreal winter. *Geophysical Research Letters*, *45*, 3605–3614. https://doi.org/10.1002/2018GL077209

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Ferro, C. (2014). Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, *140*, 1917–1923.

Fetterer, F., Knowles, K., Meier, W., & Savoie, M. (2002). *Sea ice index*. Retrieved from http://nsidc.org/data/docs/noaa/g02135_seaice_index/

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Böttinger, M., et al. (2013). Climate and carbon cycle changes from 1850 to 2100 in mpi-esm simulations for the coupled model intercomparison project phase 5. *Journal of Advances in Modeling Earth Systems*, *5*(3), 572–597. https://doi.org/10.1002/jame.20038

Gutjahr, O., Putrasahan, D., Lohmann, K., Jungclaus, J. H., von Storch, J.-S., Brüggemann, N., et al. (2019). Max planck institute earth system model (mpi-esm1.2) for the high-resolution model intercomparison project (highresmip). *Geoscientific Model Development*, *12*(7), 3241–3281. https://doi.org/10.5194/gmd-12-3241-2019

Hagemann, S., & Stacke, T. (2015). Impact of the soil hydrology scheme on simulated soil moisture memory. *Climate Dynamics*, *44*, 1731–1750. https://doi.org/10.1007/s00382-014-2221-6

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., et al. (2019). Seas5: The new ecmwf seasonal forecast system. *Geoscientific Model Development*, *12*, 1087–1117. https://doi.org/10.5194/gmd-12-1087-2019

Jungclaus, J. H., Fischer, N., Haak, H., Lohmann, K., Marotzke, J., Matei, D., et al. (2013). Characteristics of the ocean simulations in the Max Planck Institute Ocean Model (MPIOM) the ocean component of the mpi-earth system model. *Journal of Advances in Modeling Earth Systems*, *5*(2), 422–446. https://doi.org/10.1002/jame.20023

Jung, T., Miller, M., Palmer, T., Towers, P., Wedi, N., Achuthavarier, D., et al. (2012). High-resolution global climate simulations with the ECMWF Model in Project Athena: Experimental design, model climate, and seasonal forecast skill. *Journal of Climate*, *25*, 3155–3177. https://doi.org/10.1175/JCLI-D-11-00265.1

MacLachlan, C., Arribas, A., Peterson, K., Maidens, A., Fereday, D., Scaife, A., et al. (2015). Global Seasonal forecast system version 5 (GloSea5): A high resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, *141*, 1072–1084. https://doi.org/10.1002/qj.2396

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., et al. (2018). Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and its response to increasing CO_2. *Journal of Advances in Modeling Earth Systems*, *11*, 998–1038. https://doi.org/10.1029/2018MS001400

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M. A., et al. (2012). Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, *4*, 1–18.

Mogensen, K., Balmaseda, M. A., & Weaver, A. (2012). The NEMOVAR ocean data assimilation system as implemented in the ECMWF ocean analysis for system 4. (Technical Memorandum No. 66). Vol. 206. Shinfield Park, Reading: ECMWF.

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., et al. (2018). A higher-resolution version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR). *Journal of Advances in Modeling Earth Systems*, *10*(7), 1383–1413. https://doi.org/10.1029/2017MS001217

Pohlmann, H., Müller, W. A., Bittner, M., Hettrich, S., Modali, K., Pankatz, K., & Marotzke, J. (2019). Realistic quasi−biennial oscillation variability in historical and decadal hindcast simulations using CMIP6 forcing. *Geophysical Research Letters*, *46*, 14118–14125. https://doi.org/10.1029/2019GL084878

Polkova, I., Brune, S., Kadow, C., Romanova, V., Gollan, G., Baehr, J., et al. (2019). Initialization and ensemble generation for decadal climate predictions: A comparison of different methods. *Journal of Advances in Modeling Earth Systems*, *11*, 149–172. https://doi.org/10.1029/2018MS001439

Scaife, A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R., Dunstone, N., et al. (2014). Skillful long-range prediction of european and north american winters. *Geophysical Research Letters*, *41*(7), 2514–2519. https://doi.org/10.1002/2014GL059637

Scaife, A., Camp, J., Comer, R., Davis, P., Dunstone, N., Gordon, M., et al. (2019). Does increased atmospheric resolution improve seasonal climate predictions? *Atmospheric Science Letters*, *20*, 1–10. https://doi.org/10.1002/asl.922

Scaife, A., Copsey, D., Gordon, Harris, C., Hinton, T., Keeley, S., et al. (2011). Improved atlantic winter blocking in a climate model. *Geophysical Research Letters*, *38*(23), L23703. https://doi.org/10.1029/2011GL049573

Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J.-C., Hannay, C., et al. (2017). Practice and philosophy of climate model tuning across six us modeling centers. *Geoscientific Model Development*, *10*(9), 3207–3223. https://doi.org/10.5194/gmd-10-3207-2017

Siegert, S., Bellprat, O., Ménégoz, M., Stephenson, D. B., & Doblas-Reyes, F. J. (2017). Detecting Improvements in Forecast Correlation Skill: Statistical Testing and Power Analysis. *Monthly Weather Review*, *145*(2), 437–450. https://doi.org/10.1175/MWR-D-16-0037.1

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/bams-d-11-00094.1

Tibaldi, S., & Molteni, F. (1990). On the operational predictability of blocking. *Tellus A*, *42*(3), 343–365. https://doi.org/10.1034/j.1600-0870.1990.t01-2-00003.x

Wang-Chun Lai, A., Herzog, M., & Graf, H. (2017). ENSO forecasts near the spring predictability barrier and possible reasons for the recently reduced predictability. *Journal of Climate*, *31*, 815–838. https://doi.org/10.1175/JCLI-D-17-0180.1

Wilks, D. (1995). *Statistical methods in the atmospheric sciences*. San Diego: Academic Press.

Zou, H., Balmaseda, M., & Mogensen, C. (2017). The new eddy-permitting ORAP5 ocean reanalysis: description, evaluation and uncertainties in climate signals. *Climate Dynamics*, *49*, 791–811.