climate

MDPI

Article

The Added Value of Statistical Seasonal Forecasts

Folmer Krikken [1,2,3], Gertie Geertsema [2,*], Kristian Nielsen [1,4] and Alberto Troccoli [1,5]

[1] World Energy Meteorology Council, The Enterprise Centre, University Drive, Norwich NR4 7TJ, UK; folmer.krikken@climateradar.com (F.K.); kristian.nielsen@wemcouncil.org (K.N.); alberto.troccoli@wemcouncil.org (A.T.)
[2] Royal Netherlands Meteorological Institute (KNMI), P.O. Box 201, 3730 AE De Bilt, The Netherlands
[3] Climateradar, 6668 LE Randwijk, The Netherlands
[4] Underwriters Laboratories Ibérica S.L., C/ de la Caravel·la la Niña, 12, Les Corts, 08017 Barcelona, Spain
[5] School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK
* Correspondence: gertie.geertsema@knmi.nl;

Abstract: Seasonal climate predictions can assist with timely preparations for extreme episodes, such as dry or wet periods that have associated additional risks of droughts, fires and challenges for water management. Timely warnings for extreme warm summers or cold winters can aid in preparing for increased energy demand. We analyse seasonal forecasts produced by three different methods: (1) a multi-linear statistical forecasting system based on observations only; (2) a non-linear random forest model based on observations only; and (3) process-based dynamical forecast models. The statistical model is an empirical system based on multiple linear regression that is extended to include the trend over the previous 3 months in the predictors, and overfitting is further reduced by using an intermediate multiple linear regression model. This results in a significantly improved El Niño forecast skill, specifically in spring. Also, the Indian Ocean dipole (IOD) index forecast skill shows improvements, specifically in the summer and autumn months. A hybrid multi-model ensemble is constructed by combining the three forecasting methods. The different methods are used to produce seasonal forecasts (three-month means) for near-surface air temperature and monthly accumulated precipitation seasonal forecast with a lead time of one month. We find numerous regions with added value compared with multi-model ensembles based on dynamical models only. For instance, for June, July and August temperatures, added value is observed in extensive parts of both Northern and Southern America, as well as Europe.

Keywords: seasonal climate forecasts; multi-linear statistical model; non-linear statistical model; multi-model ensemble forecast

check for updates

Citation: Krikken, F.; Geertsema, G.; Nielsen, K.; Troccoli, A. The Added Value of Statistical Seasonal Forecasts. *Climate* **2024**, *12*, 83. https://doi.org/10.3390/cli12060083

Academic Editor: Nir Y. Krakauer

Received: 8 April 2024
Revised: 23 May 2024
Accepted: 29 May 2024
Published: 4 June 2024

1. Introduction

Seasonal predictions of key atmospheric variables are an important area in climate science because of their large potential value for a wide range of end users and for several applications, such as agriculture, hydrology and renewable energy (e.g., [1–3]).

Seasonal forecasts can be produced by either a statistical seasonal forecasting system or a dynamical forecasting system. Statistical methods have been used extensively (e.g., [4–6]) and are based on observed relationships between certain predictors and the forecasted atmospheric variables (predictands). Dynamical forecasting systems, on the other hand, are based on numerical models that represent the physical processes in the atmosphere, ocean and land surface and their non-linear interactions. A disadvantage of the dynamical models is that they are inherently complex and computationally expensive. Furthermore, their model output often needs further calibration due to model drift towards their preferred climate state (e.g., [7]). Statistical models do not suffer from these two issues because their relationships are based on the observations themselves. Furthermore, statistical models tend to be relatively simple and easy to interpret. Arguably, statistical models can sometimes be insufficiently sophisticated for capturing the intricate non-linear relationships among predictor variables and the predictand, particularly when longer prediction horizon times are considered (e.g., [8]).

Over the years, many comparison studies between statistical and dynamical seasonal forecasting systems have been performed. Peng et al. [9] found that for SST-forced global climate variability, dynamical and statistical models had similar forecast skill. Van Oldenborgh et al. [10] showed that for El Niño and La Niña forecasts, two dynamical ECMWF models outperformed statistical models in boreal spring and summer. However, it was found that combining the strength of both forecasting methods ultimately led to better results (e.g., [11,12]). Also, statistical models could be used to benchmark dynamical seasonal forecasts and also point to specific deficiencies in the dynamical models, such as the lack of a resolved stratosphere, needed for better forecasts of the European winter climate ([13]).

With recent advances in computer science, machine learning (ML) models have begun to be used more and more in weather forecasting (e.g., [14,15]) and seasonal forecasting. Qian et al. [16] showed that ML models provided more skilful forecasts for winter temperatures over northern and central North America than two state-of-the-art dynamical models. Another study by [17] demonstrated that a hybrid ML model using an artificial neural network outperformed the Climate Forecast System v2 in monthly ENSO indices.

Here, we analyse a suite of statistical models and assess their added information relative to a suite of dynamical models on a global scale, both in a single model and as a multi-model ensemble framework. We present an update of the relatively simple statistical forecasting system from [6], which will act as a benchmark for more advanced statistical models based on tree-based regression systems. We assess the forecast skill of the statistical models and analyse their added value relative to dynamical seasonal forecasting models in a single and multi-model forecasting setup. Nielsen et al. [18] discussed the optimization of a grand multi-model probabilistic seasonal prediction system and found that only a limited number of seasonal prediction systems was required to improve the skill. The optimal number is between three and six, with the optimal combination of models being dependent on the region and season. Based on these findings, we use a limited number of dynamical models.

The novelty of our approach is that the forecasting systems generated using different approaches are compared and combined on a global scale. The optimal combination of different systems is tested using different combinations for each grid cell, which can provide information on the performance of each system for different areas in relation to other systems.

The structure of the manuscript is as follows: Section 2 describes the data and models, including the observational data used, the different statistical models and the choices made therein, and the dynamical models that are used. The improved skill of the updated statistical model is discussed using the correlation between climate indices. Also addressed in this section is a separate note regarding verification and the use of grid-interpolated observational databases and different re-analyses datasets. Section 3 discusses the results separately for the two statistical models and then compares these methods with the dynamical models. The added value of a multi-model framework is demonstrated in Section 4. The conclusions are summarised in Section 5.

2. Data and Models

Near-surface air temperature and monthly accumulated precipitation forecasts with lead times in the order of months were investigated using a suite of statistical methods for the forecasting system, ranging from relatively simple linear regression models to more advanced tree-based regression models. Hereafter, near-surface air temperature and monthly accumulated precipitation will be abbreviated by temperature or T2M and precipitation or precip (the predictands; see Table 1).

Table 1. Overview of predictor and predictand data; see also [6] (p. 3950).

Predictand	**Acronym**	**Source** [1]
Near-surface air temperature	T2M or temperature	GHCN-CAMS [19] ERSSTV5 [20]
Monthly accumulated precipitation	Precip or precipitation	GPCC [21]
Predictor type	**Predictor**	**Source**
Climate index predictor	NINO34 NINO12 WWV PDO AMO IOD	Based on ERSSTV5 [2] POAMA PEODAS [22] Based on ERSSTV5 ERSSTV5 [23] Based on ERSSTV5
Local predictors	Precipitation Persistence	GPCC [21] Same as predictand
Long-term trend	CO2EQ	CO2-equivalent concentrations [24]

[1] Observational gridded databases. [2] Normalized relative to 1981–2010.

2.1. Data

Statistical forecasts are based on observed relationships between certain predictors and the forecasted atmospheric variables. Hence, we need reliable observational gridded products that: (i) cover a long time period, (ii) are frequently updated and (iii) have a high enough spatial resolution. For temperature, we used the Global Historical Climatology Network version 2 and the Climate Anomaly Monitoring System (GHCN-CAMS) dataset ([19]; 0.5° resolution). Over the oceans, we used the Extended Reconstructed Sea Surface Temperature V5 (ERSSTV5) ([20]; 2° horizontal resolution) dataset. This combination of SST and T2M is hereafter labelled as T2M. For precipitation, we used the Global Precipitation Climatology Centre (GPCC) dataset ([21]; 1° hor. resolution). For all products, we used monthly data. The predictors we used (Table 1) can be divided into local predictors (spatio-temporal data), large-scale climate indices and the CO2-equivalent forcing (CO2EQ) as a predictor for the long-term trend.

As local predictors, we used persistence. The previous three-month average was used to predict the following upcoming three months: for example, OND_{2020} was used as a predictor for FMA_{2021}. In other words: OND_{2020} is the temperature or the precipitation averaged over the months of October, November and December and used as a predictor for the three-month averaged temperature or precipitation for the following February, March and April. Because of the relation between warm periods following from precipitation deficits ([25]), we also used the previous 3-month accumulated precipitation as a predictor for the next 3-month temperature.

For the large-scale climate indices, we used the El Niño Southern Oscillation (ENSO), Pacific Decadal Oscillation (PDO), Atlantic Multidecadal Oscillation (AMO), Indian Ocean dipole (IOD), North Atlantic Oscillation (NAO) and other predictors.

In the predictor choices, we closely followed [6], who made a selection based on physical principles and processes so as to benefit from the predictive power from the long-term trend associated with the CO2EQ and at the same time minimise the risk of overfitting by adding as few additional predictors as possible. We refer to [6] for a more in-depth analysis of the predictive power of these climate indices. In Table 1, the predictors are listed that are found to significantly add value to the predictions, with the datasets used. The time period ranged from 1961 to 2021. The data were re-gridded to a 1° × 1° grid. ENSO was quantified using multiple indices, namely the NINO34 (average SST over 5N–5S and 170W–120W) and NINO12 (average SST over 0–10S and 90W–80W) index and the warm water volume (WWV). All SST-based indices are based on ERSST V5.

2.2. Statistical Models

The different statistical models were constructed in a similar manner. The models forecast the next 3-month mean, i.e., a forecast initiated in January will produce a forecast for FMA. For each grid point, a separate model was fitted. The predictor and predictand data were first detrended based on a linear regression with CO2EQ in order to have a stationary dataset. All models have a global coverage and a horizontal resolution of 1 degree. Forecasts were assessed using the Leave-1-year-out cross-validation over the period of 1961 to 2020.

2.2.1. Multiple Linear Regression (MLR)

The MLR model is an updated version of the relatively simple statistical forecasting system from [6]. It forecasts the next 3-month average based on the previous 3-month mean predictors. As an example, a forecast of T2M issued in January is based on OND predictor data and predicts FMA.

In this work, the [6] model was extended and improved by incorporating second-order information as predictors, namely the 3-month linear trend. This was deemed necessary following an analysis of forecast errors, which identified that situations with a change of sign in the NINO34 index during the period leading up to the forecast resulted in large forecast errors. This issue arises because the 3-month average of the predictors (the NINO34 index in this case) may obscure important variations within the three-month period. To avoid overfitting by introducing too many predictors, we have used an intermediate MLR model to predict the future state of each climate index based on its previous 3-month mean and 3-month trend. These future states of the climate indices were then used as predictors in the actual MLR model. For the NINO34 index, this greatly reduces the spring predictability barrier, and it increases the autumn IOD forecast predictability (Figure 1).

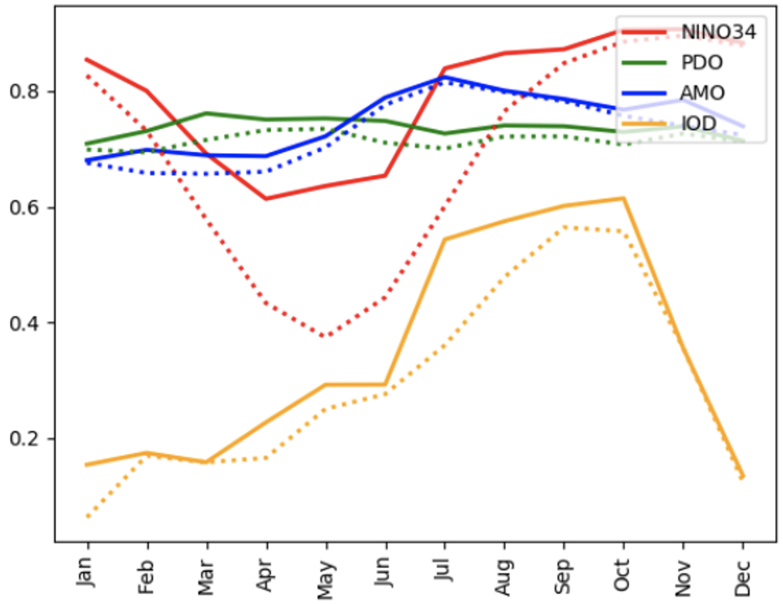


Figure 1. Lagged correlation (dashed line) and the intermediate MLR model (solid line) for the climate indices. The dashed lines represent the lagged correlation between the 3-month mean and the 3-month mean 3 months ahead; e.g., January shows the correlation between the OND average and the next FMA. The solid lines show the correlation between the forecasted climate index and the observed climate index. The forecasted climate index is based on a multiple linear regression model, with the previous 3-month mean and previous 3-month trend being used as predictors.

For the local predictors (e.g., persistence), the intermediate MLR step is skipped, and the 3-month trend is added as an extra predictor in the actual MLR model.

Another update to the original forecasting system is to have a stricter predictor selection routine. In [6], only predictors with a significant correlation ($p < 0.05$) with the predictand were added. Though this removes non-relevant predictors, it does not account for co-variability between predictors, which can lead to increased overfitting of the model. This can be an issue for instance in the NINO34 region, where we expect a strong correlation between the NINO34 index and SST (persistence predictor). To overcome this, we implemented a stepwise predictor selection routine. In step 1, we compute the correlation between all predictors and the predictand. The predictor with the highest significant correlation ($p < 0.05$) is then chosen as the first predictor in the MLR model. In step 2, we remove the linear relation between the chosen predictor and the predictand from the predictand time series, creating a residual time series (PR_{RES}). In step 3, we again compute the correlation between the predictors and the predictand times series, but now with the remaining predictors and with PR_{RES}. The predictor with the highest significant correlation is then chosen as the second predictor. These steps are continued until there is no significant correlation between a predictor and PR_{RES}. This procedure was performed for every grid point individually. With the leave-1-year-out cross-validation, the residuals can be calculated for each timestep for the period of 1961–2020 minus the one year that was used for the cross-validation.

The ensemble was calculated by randomly sampling 50 values from the residuals (forecast error) of the model fit and adding these to the deterministic forecast. If there is poor predictability, the errors will be large; thus the ensemble spread will be relatively large. If there is good predictability, the errors will be small; thus, the ensemble spread will be relatively small.

The advantages of using the relatively simple MLR method is that we can identify the individual contribution of each predictor and it works well on relatively small sample sizes. The disadvantages of using MLR is that it assumes a normal distribution and homoscedasticity, is sensitive to outliers and assumes a linear relationship.

Note that besides MLR, we also tested LASSO and ridge regression but found no improvement relative to MLR; hence, we did not proceed with these methods.

2.2.2. Random Forest Regression (RFR)

Random forest regression (RFR) is a tree-based ensemble regression model, which is a popular machine learning tool used for many different forecasting problems and research fields. We used the random forest regressor of the Python module scikit-learn (see [26]). Random forests are constructed by individual decision trees. A basic decision tree model can make very accurate predictions on the data it was trained on. However, it generally leads to very bad results on new (testing) data. To circumvent this, multiple decision trees (i.e., a forest) can be built, where each decision tree uses a bootstrapped sample of the original training data and then takes the average of all these trees. This method is known as bagging, i.e., taking the aggregate of all the different trees based on bootstrapped training data. The main advantage of this method, implemented by the RFR model, is that it performs much better on new data, thus leading to better generalization.

With RFR, there are several tuning parameters. Given that the model is fit for each grid point, the optimal parameter combination will differ per grid point. A sensitivity analysis for the different parameters showed that the RFR model was mostly sensitive to the maximum depth of the decision trees. Hence, a parameter selection routine on the maximum depth ranging from 1 to 7 was performed for each grid point separately to establish the best possible parameter setting. Generally, in regions with low forecast skill, the maximum depth is kept smaller to avoid overfitting, whereas in regions with larger predictability, the maximum depth is higher to allow for more complex trees. Other relevant parameters were set to a fixed value. The number of trees in the forest were set to 50, and the maximum number of predictors was set to 3 in order to prevent overfitting.

RFR models tend to be 'data hungry' [27], i.e., they require a large training sample in order to yield stable models. The data used in this study cover the period from 1961 to 2020;

thus there are about 60 samples if the model is fitted for each month individually (therefore retaining the annual cycle) and about 700 samples if the model is fitted using all months together. Ideally, the model should be fitted for each individual month because the predictor–predictand relations can strongly differ for different months. However, initial tests pointed to stronger overfitting using only 60 samples relative to using the full dataset. This is why we constructed two RFR models, one that fits a model for each month individually (RFR-M), and one model that uses the full sample (all months together, RFR-Y).

Note that in the RFR-Y model we normalise the data by dividing the predictand data by its monthly standard deviation prior to the model fit and then multiplying the forecast by its monthly standard deviation to yield the actual value of the variable considered. This is performed because variability is month-dependent, especially for precipitation, which can lead to unrealistic high variability over months with relatively low variability.

We increased the number of potential predictors in the model by adding more permutations of the predictor data to the model. Besides the 3-month mean and 3-month trend, as performed in the MLR model, we also added the 5-month mean and 1-month values at lags of 1, 3 and 5 months. We tested multiple predictor selection routines, but found no clear reduction in forecast error, and we greatly increased the model run time. Hence, we did not use a predictor selection routine for the RFR models.

We constructed probabilistic forecasts by taking into account all trees of the random forest and not just the average or median value over these trees, as is usually done. Given that the number of estimators (trees in the forest) is 50, we have an ensemble size of 50. The advantages of RFR relative to MLR is that it assumes no distribution and can handle non-linear relationships and heteroscedasticity.

We also tested several other tree-based regression models such as the gradient boosting method and regression-enhanced random forest, but these methods showed no clear improvement of the standard RFR; hence, we did not proceed with these methods.

Figure 2 provides a schematic diagram of the time schedule of the predictands used by both statistical models.

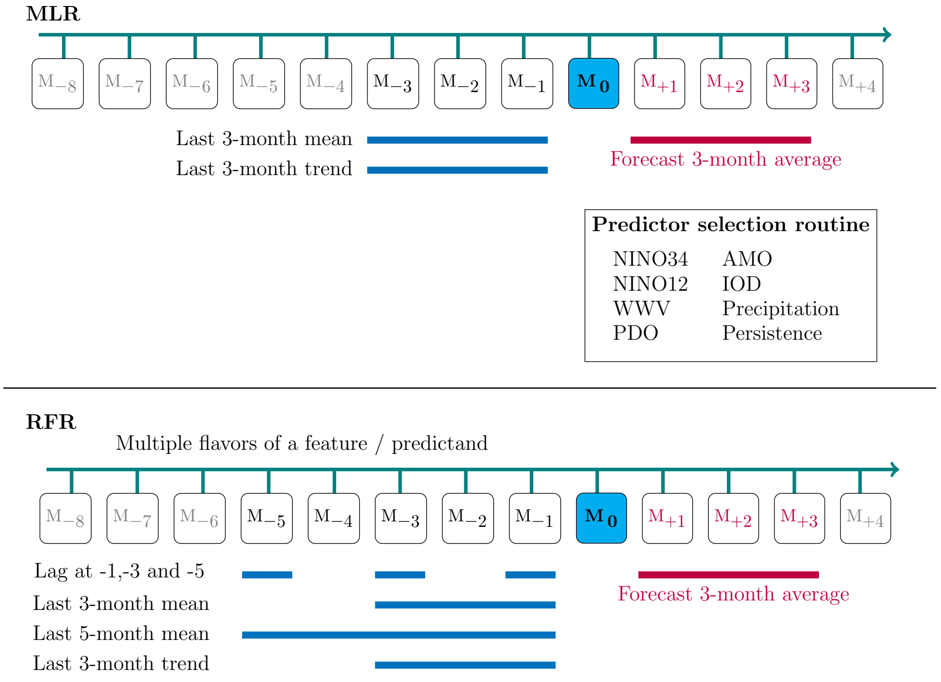

MLR
M_{-8} M_{-7} M_{-6} M_{-5} M_{-4} M_{-3} M_{-2} M_{-1} M_0 M_{+1} M_{+2} M_{+3} M_{+4}
Last 3-month mean
Last 3-month trend
Forecast 3-month average
Predictor selection routine
NINO34
AMO
NINO12
IOD
WWV
Precipitation
PDO
Persistence
RFR
Multiple flavors of a feature / predictand
M_{-8} M_{-7} M_{-6} M_{-5} M_{-4} M_{-3} M_{-2} M_{-1} M_0 M_{+1} M_{+2} M_{+3} M_{+4}
Lag at -1,-3 and -5
Last 3-month mean
Last 5-month mean
Last 3-month trend
Forecast 3-month average

Figure 2. Schematic diagram of the prediction time schedule. The selection routine selects predictors from the set of predictors for each predictand and for each grid cell separately.

2.3. Dynamical Models

The dynamical model forecasts used here were retrieved by the SECLI-FIRM team at a 1° × 1° regular lat/lon grid. A summary of the resulting database is given in Table 2. The analysis here was conducted for the years of 1993–2016 as this period of 24 years is covered by all models. More details on the retrieval can be found in [18] (p. 9).

Table 2. List of dynamical seasonal forecasting models, taken in hindcast mode (hence with lower ensemble members than for forecast, and in some cases different initialization datasets).

Name	Producer	Atmospheric Model Resolution	Ocean Model	Members	Initialization Atmos	Initialization Ocean
CANI	Meteorological Service of Canada (MSC), Canada	CanAM4 T63, corresponding to a 128 × 64 Gaussian grid Res: ∼2.8°lat/lon	CanOM4 Res: ∼1.4°(lon), ∼0.94°(lat)	10	CMC	ORAP5 ocean reanalysis
CCSM	University of Miami (RSMAS), USA	CAM4 0.9°× 1.25°	CCSM POP2 g1v6 ∼1° × 1°	10	CFSR	OISST
CMCC	CMCC, Centro Euro-Mediterraneo per i Cambiamenti Climatici, Italy	CESM 1.2—CLM5.3 Res.: 0.5 lat/lon L46	NEMO v3.4 Res.: ORCA0.25 L50	40	ERA5	C-GLORS Global Ocean 3D-VAR
DWD	Der Deutsche Wetterdienst (DWD), Germany	ECHAM 6.3.05—JSBACH v3.20 Res.: T127 L95	MPIOM 1.6.3 Res.: TP04 L40	30	ERA5	ORAS5
ECMWF	European Centre for Medium-Range Weather Forecasts (ECMWF), EU	IFS Cycle 43r1 Res.: Tco319 L91	NEMO v3.4 Res.: ORCA0.25	25	ERA-Interim	ORAS5
GEMN	Meteorological Service of Canada (MSC), Canada	GEM 4.8-LTS.13 Res: ∼110 × 110 km	NEMO 3.6 ORCA Res: ORCA1 ∼1°lat/lon	10	ERA-Interim	ORAP5 ocean reanalysis
GFDL	The Geophysical Fluid Dynamics Laboratory (GFDL) , USA	AM 4.0 1° × 1° , L33	MOMv6: ∼1° × 1° , tropical refinement to 0.3°	15	CFSR	OISST v2
JMA	Japan Meteorological Agency (JMA) / Meteorological Research Institute (MRI)	JMA-GSM Res: TL159 (110 km)	MRI.COM v3 Res: 1°× 0.3°/ 0.5°on a tripolar grid	10	JRA-55	MOVE / MRI.COM-G2
MF	Meteo France. France	ARPEGE v6.2 - SURFEX v8.1 Res.: T359 L91	NEMO v3.4 Res.: ORCA1 L75	25	ERA-Interim	GLORYS2V2
NCEP	National Center for Environmental Prediction, USA	GFS Res: T128 (∼1° × 1°)	GFDL MOM4 Res: ∼0.5°lat/lon	24	CFSR	CFSR
UKMO	UK-Met Office Met Office, United Kingdom	Unified Model (UM)—Global Atmosphere 6.0—JULES 6.0 Res.: N216 L85	NEMO v3.4—Global Ocean 5.0 Res.: ORCA0.25 L75	28	ERA-Interim	GS-OSIA

In order to assess the relative accuracy of statistical and dynamical models, we compared the statistical models to 11 dynamical seasonal forecasts (listed in Table 2). The seasonal forecasts were all bias corrected by simply subtracting the mean bias, computed as the difference between the ensemble mean in the hindcast mean and observations. This was performed for each forecasted period separately. As an example, all forecasts valid for JJA have the same correction factor.

2.4. Verification

Several options have been considered in terms of reference data for the verification of seasonal climate forecasts ranging from climatology based on reanalyses products to observational gridded datasets. For verification on a global scale, the best option is not obvious a priori.

Despite significant improvements in observational products in recent times, considerable observational uncertainty still remains for certain regions. Different observational products can differ considerably, making the choice of which product to use for verification non-trivial. Figure 3 shows the observational uncertainty, quantified by the disagreement (standard deviation) between multiple observational and reanalysis products (listed in Table 3). Here, we first compute the standard deviation (STD) per time step (monthly average) between the different observational products and then average this STD over all months and years. Generally speaking, there is a larger uncertainty around regions with less observations and in areas with complex terrain. However, there are many regions with a standard deviation of around 0.5 °C, indicating a spread of around 1.5 °C 3 sigma).

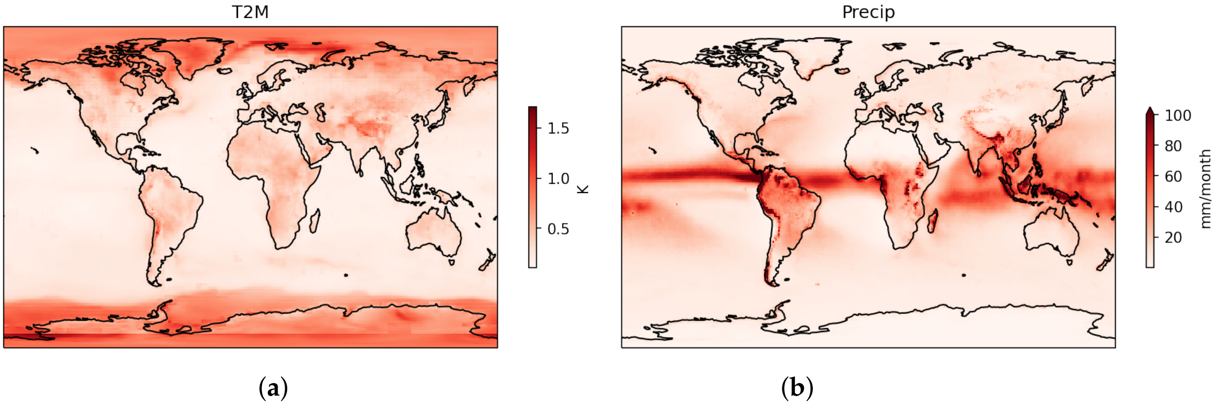


Figure 3. The average standard deviation between all products is used to quantify the observational uncertainty for (**a**) the near-surface temperature (T2M) in Kelvin and (**b**) precipitation (Precip) in mm/month. The standard deviation is averaged over the full time period (1980–2016).

Table 3. List of observational and reanalysis products used.

Product	T2M	Precip.	Resolution	Reference
ERA5	x	x	0.25°	[28]
JRA-55	x	x	1.25°	[29]
MERRA-2	x	x	0.5°	[30]
ERSST V5	x	-	2°	[20]
CRUTEM	x	-	5°	[31]
GPCC	-	x	0.5°	[21]
GISTEMP	x	-	2°	[32]

For precipitation, the STD can be as large as 100 mm/month or even larger (Figure 3). Over the oceans, the STD is only calculated between the different reanalysis products, showing that the uncertainty there is of the same order.

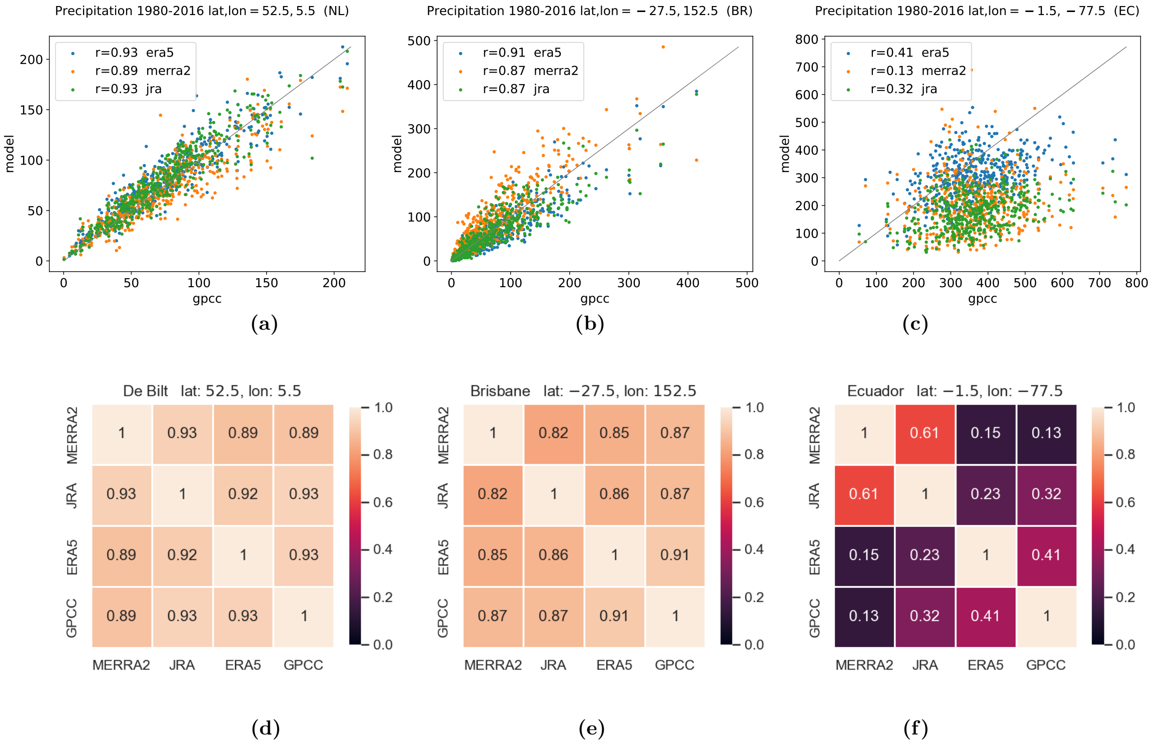
These large STDs occur specifically around the equator, where the climatological monthly averages range from 100 mm/month to several hundreds of mm/month.

For the observational gridded precipitation database GPCC, [21,33] discussed the different sources of errors in the rain gauges-based gridded monthly precipitation database. Schneider et al. [21] estimated the relative sampling error with respect to the true area mean for different numbers of stations per grid to be between ±7 to 40% for five sites and between ±5 and 20% if there are ten sites inside the grid box.

The GPCC dataset provides precipitation data along with the number of stations per grid. We used this information to select two grid locations with many stations in a grid box and one location with only a few stations in a grid box. The three grid locations are taken within The Netherlands, the greater Brisbane region and Ecuador, respectively. The number of stations per grid was, for a randomly chosen month, 60, 96 and 2, respectively. We compared the GPCC precipitation with the reanalysis products ERA5, MERRA-2 and JAM (see Table 3). The reanalysis data were interpolated to the same 1° × 1° grid.

Figure 4 shows the correlation between the reanalysed products with GPCC data as scatter plots; also shown are correlation matrices for the same locations.

The correlation between the reanalysed products with the observed values is low for the grid box with only a few measurements, as can be expected. However, the correlation matrices show that the correlation between the reanalysed products themselves is low too. Hence, from these correlation matrices, it is difficult to conclude which product is the best reference dataset.



(a) (b) (c)

(d) (e) (f)

Figure 4. (**a**–**c**): The top row shows the monthly accumulated precipitation for the models as a function of the observed precipitation for ERA5 (blue), Merra-2 (orange) and JRA (green) for areas near (**a**) De Bilt, The Netherlands, where the maximum monthly precipitation is around 250 mm/month; (**b**) Brisbane, Australia, with a maximum monthly precipitation around 500 mm/month; and (**c**) Ecuador, with a maximum monthly precipitation around 800 mm/month. Note that the axes reflect the differences in precipitation. (**d**–**f**): The bottom row contains the correlation matrices for the same models and locations: (**d**) De Bilt, The Netherlands, (**e**) Brisbane, Australia and (**f**) Ecuador.

The statistical models are all biased towards their own reference product, whilst the dynamical models are also all biased towards their respective observational product used for their initialization. Hence, which model performs better will strongly depend on which observational product is used as a reference. To circumvent this potential bias, we used the average of multiple observational and reanalysis products (ENS_OBS) as a reference dataset to evaluate the forecast skill.

3. Results

The forecast skill is quantified by the continuous ranked probability score (CRPS), which is an often-used measure for probabilistic forecasts. It quantifies the error based on the quadratic measure of the difference between the forecast cumulative density function and the observed value. We use the CRPS skill score variant (CRPSS) by directly comparing the CRPS of the forecast to a reference forecast (CRPSS = 1 − [$\text{CRPS}_{forecast}$ / $\text{CRPS}_{reference}$]). Positive values indicate that the forecast outperforms a reference forecast.

3.1. Multiple Linear Regression (MLR)

First, we look at the forecast skill of the relatively simple MLR model. We use a climatological forecast as reference, which is constructed by randomly sampling 50 values (same size as MLR and RFR ensemble) from the climatology with leave-1-year-out cross-validation. The skill score is based on data from 1980 to 2016. Only significant values ($p < 0.05$) are plotted, based on a bootstrapped sample. Positive CRPSS values indicate that the MLR forecast performs better than the climatological forecast. Figure 5 shows the CRPSS of T2M and PRECIP for the JJA and DJF forecasts, initiated in May and November, respectively. Positive CRPSS values indicate that the MLR forecast performs better than the climatological forecast; these are the red areas in Figure 5.

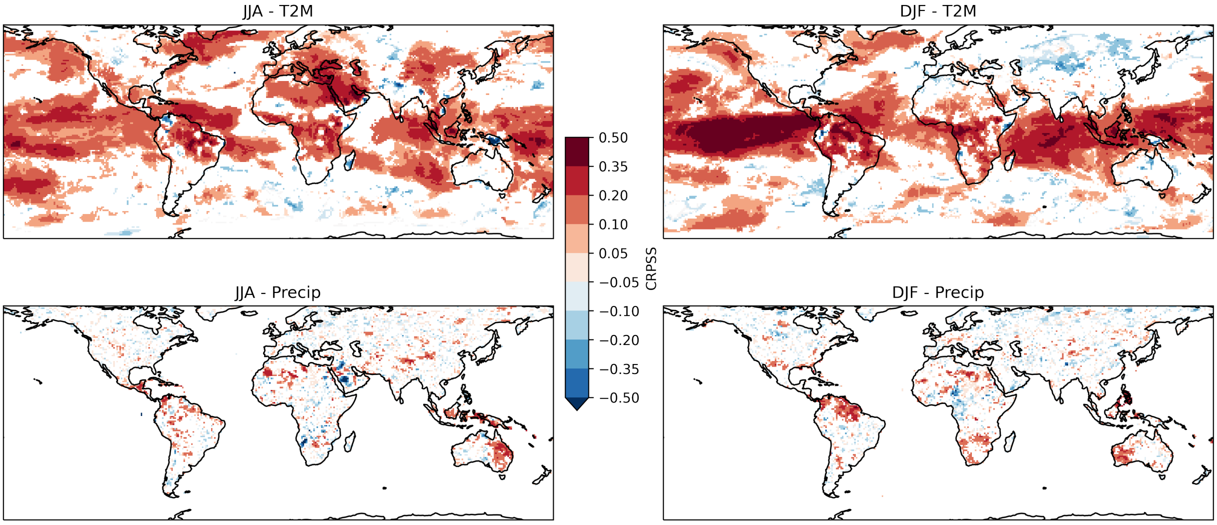


Figure 5. Forecast skill (CRPSS) with a climatological forecast as a reference for the MLR model for T2M and Precip and for JJA and DJF. Only significant values ($p < 0.05$) are plotted, based on a bootstrapped sample. The skill score is based on data from 1980 to 2016. Positive CRPSS values indicate that the MLR forecast performs better than the climatological forecast. (**Top left**) T2M CRPSS for JJA, (**top right**) T2M CRPSS for DJF, (**bottom left**) Precip CRPSS for JJA, (**bottom right**) Precip CRPSS for DJF.

It is clear from the T2M figures that forecast skill is largest in the tropical regions, owing to the large influence of ENSO. Also, skill over the ocean is generally larger because of the stronger persistence of anomalies. In JJA, we also find some skill over Europe, which is mainly related to the long-term trend.

For precipitation, the forecast skill is generally much lower. The teleconnections with ENSO do provide some skill over the northern part of South America, Australia and southern part of North America. Negative values (worse than a climatological forecast) for both T2M and precipitation are to some extent caused by overfitting, but mostly due to differences in between GHCN (used as the observational estimate in the model fit) and ENS_OBS. For an extensive evaluation of the added benefit of the individual predictors, we refer to [6].

3.2. Random Forest Regression (RFR)

Next, we assess the forecast skill of the RFR-M model, where the MLR forecast is used as reference (Figure 6). Hence, red values indicate that the RFR-M model outperforms the MLR model, whereas blue values indicate that the MLR model outperforms the RFR-M model. RFR-M outperforms the MLR forecasts over several regions, but there are also many regions where MLR outperforms the RFR-M model. There are distinct regions where one model performs better than the other, such as Russia for MLR and the Western Pacific for RFR-M. For precipitation, MLR outperforms the RFR-M model in most areas, with the exception of an area in Southern America for DJF and Northern Africa in JJA.

When comparing RFR-Y with MLR (Figure 7), we again find some regions where RFR-Y outperforms MLR. However, there are considerably more regions where MLR outperforms RFR-Y.

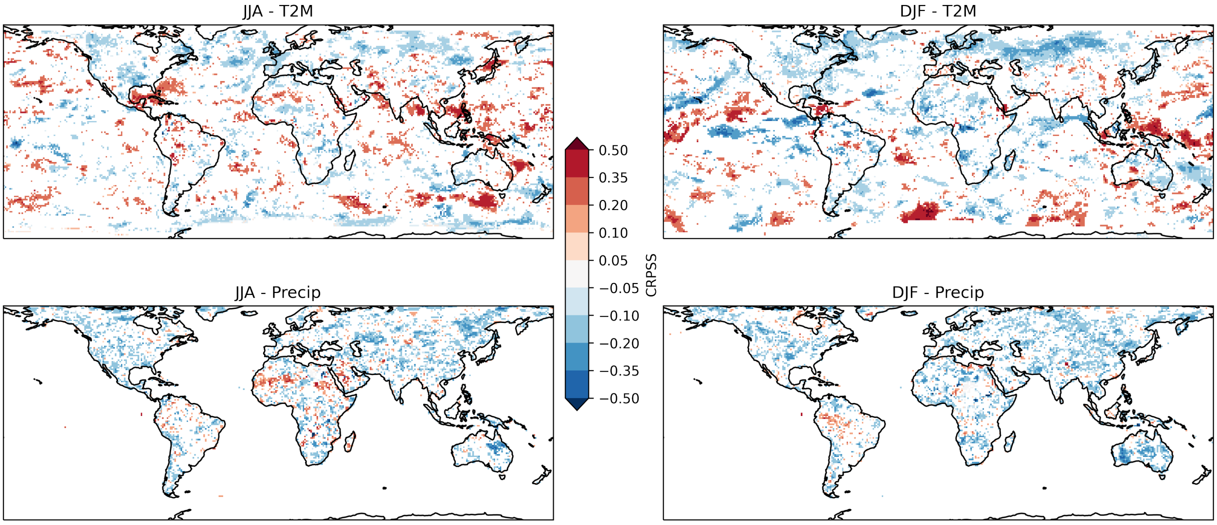


Figure 6. As in Figure 5, but for RFR-M with MLR as the reference forecast.

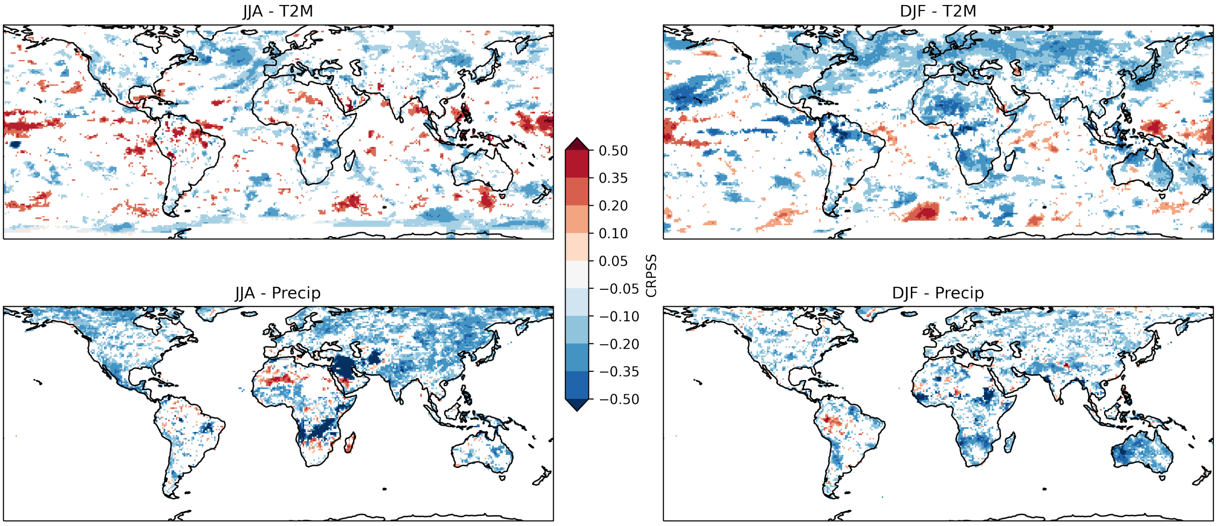


Figure 7. As in Figure 6, but for RFR-Y instead of RFR-M.

The RFR-Y model fails to reproduce some of the variability in the northern regions, whereas MLR and RFR-M are instead capable of reproducing some of the variability. It seems that for many regions, the seasonal relation between predictor and predictand differs considerably. This leads to worse results when pooling all months together and missing part of the variability that is correctly forecasted otherwise. For precipitation, the MLR model seems to outperform RFR-Y in almost all regions.

The results indicate that using more advanced models does not necessarily lead to better results. RFR models generally need a large training set in order to create stable models, and it seems that 60 years of data is not a large enough sample to really outperform MLR. By pooling all months together (RFR-Y), the sample size is largely increased, but at the cost of losing the individual predictor–predictand relations on a monthly basis. Especially for ENSO, which is phase-locked to the seasonal cycle, the relations strongly differ throughout the year, making RFR-Y less skilful than RFR-M.

3.3. Statistical versus Dynamical Models

The advantage of statistical models relative to dynamical models (DYNs) is the low computation costs and an easier understanding of the sources of predictability, either through the regression coefficients (MLR) or features importance (RFR). However, it is also important to know whether statistical models provide added information relative to dynamical models. All of the analysis in this section was performed on data ranging from 1993 to 2016 because of the availability of the hindcasts of the dynamical forecasts.

In order to assess the added value of the statistical models, we compared them with a set of dynamical forecasts (listed in Table 1). We calculated the CRPS of each model (ENS_OBS as observational estimate) and selected the best performing model (lowest CRPS) per grid point (Figure 8). The labels in Figure 8 denote the type of model which performed best. It is clear that for most regions, one of the dynamical models performed best. However, especially in the JJA, both for T2M and precipitation forecasts, there are still many regions where either an RFR model or MLR model performed best, such as parts of central and North America and the UK.

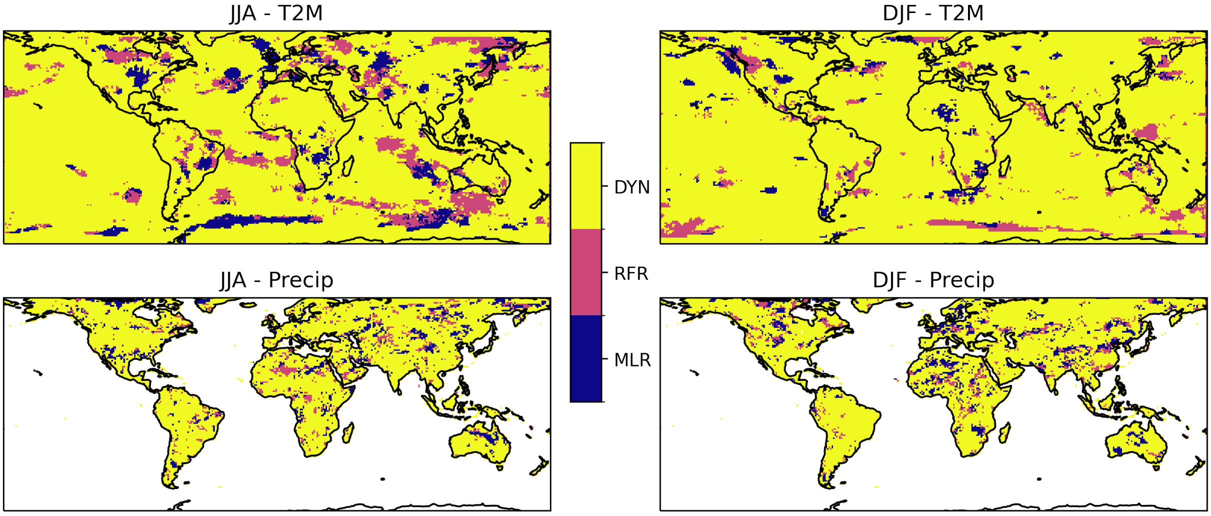


Figure 8. Best scoring model type per grid point based on the minimum CRPS.

4. Discussion: Added Value in a Multi-Model Framework

Though the results in Figure 8 indicate that dynamical models provide the best individual forecast in general, this does not automatically indicate that there is no added value in using statistical forecasts. In general, a multi-model combination of seasonal forecasts tends to outperform single seasonal forecasts ([34]; see also previous sections in this report). Hence, to fully assess the added value of statistical models, we constructed a multi-model forecast of five models out of every possible combination of dynamical and statistical models available. For the number of models in this analysis, this amounted to 2002 individual multi-model combinations. This number was chosen because the average number of models in the optimal multi-model combination was estimated at 4.5 models for the two variables over the domains [18]. In order to assess statistical significance, we constructed 100 different samples through resampling with replacement (bootstrapping, n = 100). For each sample, we selected the model with the lowest RMSE, after which we counted how many times a certain model was present in one of the 100 resampled samples. In order to test whether a model was chosen more often than due to random chance, we used a binomial significance test ($p < 0.05$). From the list of significant models, we counted how many times a statistical model was present, which indicates added value for statistical models in a multi-model framework. This analysis was performed for each grid point individually as seen in Figure 9. Note that, due to computational constraints, we limited this analysis to deterministic skill metrics (i.e., ensemble averages). It is clear that compared with Figure 8, there are many more regions where a statistical model provided additional information in a multi-model framework. For JJA and T2M, there are a number of regions (North America, South America and Europe) where at least one of the three models are statistical models. For precipitation, the results are a bit more scattered but do point to the same conclusion that, in a multi-model framework, there is regionally added value by combining statistical with dynamical models.

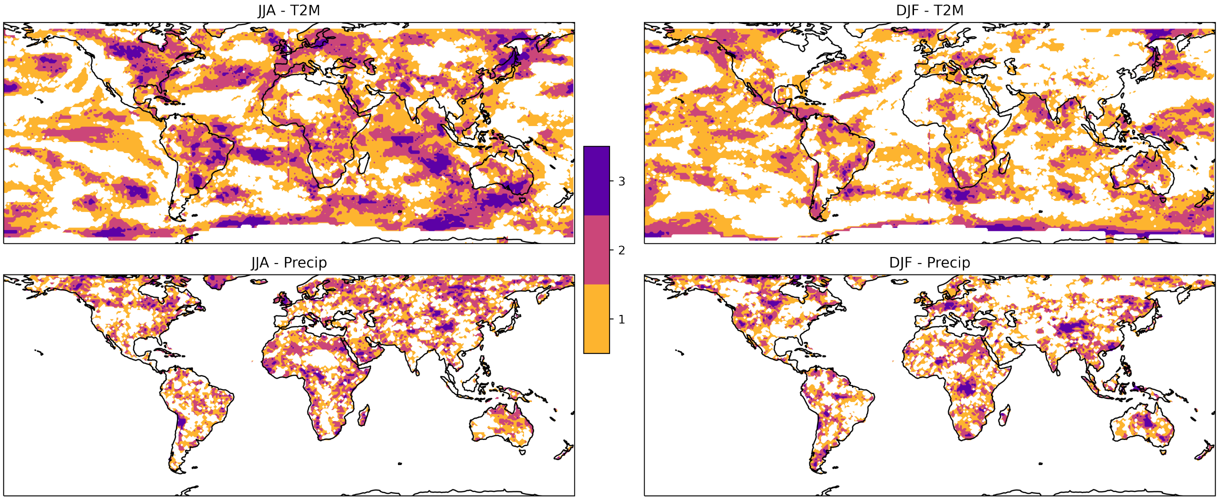


Figure 9. Added value of statistical models in a multi-model framework. The values indicate how many statistical models were listed in the best performing statistically significant multi-model combination of 5 models. Statistical significance was acquired through bootstrapping (n = 100).

Although Figure 9 shows there are many areas where statistical models can provide added value, it is difficult to assess the actual magnitude of the added value. In order to quantify this, we selected the best model combination with at least one statistical model included and the best model combination with no statistical model included and calculated the percentage change in RMSE. The results are shown in Figure 10. Blue values indicate that adding the statistical model to the dynamical model improves the result (reduction in RMSE). For instance, for the JJA temperature forecast, there is added value in large parts of Northern and Southern America and Europe. For Europe and North America, the forecast error reduces by up to 10–20%. There are also areas where the forecast is deteriorated by adding a statistical model, which are mostly situated over the ocean. Further research is needed to better understand the sources of this increased forecast skill due to adding a statistical model.

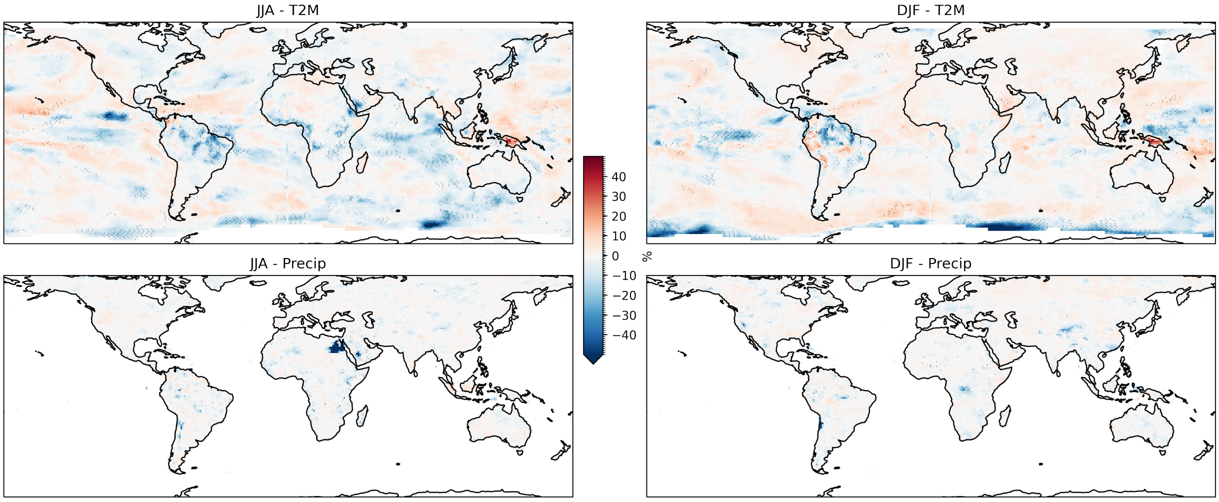


Figure 10. Change in RMSE for the best multi-model combination with and without a statistical model (using 5 models in the combination). Blue values indicate that adding a statistical model lowered the RMSE. The values are based on the average of the bootstrapped sample (n = 100). Dots indicate where 90% of the resampled samples agree with the sign of the change.

Note that there is an inherent uncertainty in the findings due to the relatively short hindcast period and observational uncertainty. Also, we used a fixed value of five models in the multi-model ensemble, while locally, this number might be more or less if an objective method was used to select the best model combination. Hence, the fixed value of five could artificially increase the value of the dynamical or statistical models. More work is needed to better quantify these uncertainties. The setup with a three-model combination yielded similar results. Note that when using a five-model combination, there are a total of 2002 possible multi-model combinations. Out of these, there are 462 combinations with no statistical models and 1540 combinations with at least one statistical model. In order to create an unbiased comparison, Figure 10 is based on a randomly selected subset (462 out of the 1540 combinations). From these results, it is clear that when constructing a multi-model, in this case with a size of five, from eleven dynamical systems and three statistical models, the latter models seem to often be needed to improve the skill of the forecast.

Figure 10 shows the results when the multi-model combination is tested against OBS_ENS (the average of observational and reanalysis products). To test the sensitivity to the reference product, we also tested the performance of the multi-model combination against individual observational and reanalysis reference products. We found only small sensitivity to which product is used as reference, indicating that the result shown in Figure 10 is a robust result. This indicates that, in general, there is a large added value of statistical models in a multi-model framework.

5. Conclusions

In this paper, we analysed the skill of relatively simple and more advanced statistical seasonal climate forecasting systems and assessed their added value relative to dynamical seasonal forecasting systems, both as single-system forecasts and in a multi-model context.

The relatively simple seasonal forecast is based on multiple linear regression. In this paper, the model is extended to include second-order (i.e., differential) information in the predictors, and overfitting is further reduced by using an intermediate multiple linear regression model. This update results in a significantly improved NINO34 index forecast skill, specifically in spring. The forecasts display skill in many regions, particularly for near-surface air temperature. It has skilful forecasts in the tropical regions, where there are strong teleconnections with large-scale climate indices such as NINO34. The persistence of anomalies and the long-term trend is also a large source of forecast skill.

Machine learning models (RFR) improve the forecasts locally, but the small sample size hampers their forecast skill. By using the full sample (all months pooled together), the forecasts become more stable (less overfitting). This improves the forecast in certain regions but mostly reduces forecast skill in other regions because it loses the individual predictand–predictor relations that differ throughout the year. Hence, we find no ‘best’ model for all grid points, but rather an ensemble of statistical empirical models whose skill depends on the region considered.

When comparing the statistical models to a suite of dynamical models, we find that, in general, the best individual model is one of the dynamical models, though the specific model varies depending on the area. There are some regions where the best forecast skill is obtained by a statistical model, but this is rather limited. In a multi-model framework, however, there are numerous regions where the multi-model average forecasts are improved by a combination of statistical and dynamical models instead of only using a combination of dynamical models. For instance, for the JJA temperature forecast, there is added value in large parts of Northern and Southern America and Europe. For Europe and North America, the forecast error reduces by up to 10–20%.

Our findings are based on a relative short hindcast period and a fixed number of five models in the multi-model ensemble. The results can be further improved by extending the hindcast period and optimizing the number of models in the multi-model ensemble per grid box by an objective method.

The global assessment of the hybrid multi-model ensemble combining dynamical and statistical models demonstrates the added value using an observational dataset constructed from a combination of reanalysis and observational products. More sophisticated methods for analysing the added value would probably yield even more added value, for example focussing on regions where observationally based datasets and the different reanalysis datasets agree. Identifying seasons (months) and regions where the statistical models add value can help to explore the full benefit of the hybrid multi-model ensemble for seasonal climate forecasts. Further development could target specific areas where the inclusion of the statistical models already shows added value.

Given that marginal improvements in seasonal forecasts of precipitation and temperature are already very useful for sectors such as energy or agriculture, these results highlight the need for adding statistical models to multi-model ensemble seasonal forecasts.

Author Contributions: Conceptualization: F.K., G.G., K.N. and A.T.; Methodology: F.K., G.G., K.N. and A.T.; Formal analysis, F.K. and G.G.; Resources and data curation,: K.N., A.T., F.K. and G.G.; Writing—original draft preparation, F.K.; Writing—review and editing, G.G., F.K., A.T. and K.N.; Visualization, F.K. and G.G.; Funding acquisition: A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was part of the SECLI-FIRM project, which was funded by the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement 776868.

Data Availability Statement: Data are available in publicly accessible repositories mentioned in Tables 2 and 3. The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: Author Folmer Krikken was employed by the company Climateradar. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

References

1. Rodriguez, D.; de Voil, P.; Hudson, D.; Brown, J.N.; Hayman, P.; Marrou, H.; Meinke, H. Predicting optimum crop designs using crop models and seasonal climate forecasts. *Sci. Rep.* **2018**, *8*, 2231. [CrossRef] [PubMed]
2. Demirel, M.C.; Booij, M.J.; Hoekstra, A.Y. The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models. *Hydrol. Earth Syst. Sci.* **2015**, *19*, 275–291. [CrossRef]
3. Torralba, V.; Doblas-Reyes, F.J.; MacLeod, D.; Christel, I.; Davis, M. Seasonal Climate Prediction: A New Source of Information for the Management of Wind Energy Resources. *J. Appl. Meteorol. Climatol.* **2017**, *56*, 1231–1247. [CrossRef]
4. Barnston, A.G.; Glantz, M.H.; He, Y. Predictive Skill of Statistical and Dynamical Climate Models in SST Forecasts during the 1997–98 El Niño Episode and the 1998 La Niña Onset. *Bull. Am. Meteorol. Soc.* **1999**, *80*, 217–244. [CrossRef]
5. Landsea, C.W.; Knaff, J.A. How Much Skill Was There in Forecasting the Very Strong 1997–98 El Niño? *Bull. Am. Meteorol. Soc.* **2000**, *81*, 2107–2120. [CrossRef]
6. Eden, J.M.; van Oldenborgh, G.J.; Hawkins, E.; Suckling, E.B. A global empirical system for probabilistic seasonal climate prediction. *Geosci. Model Dev.* **2015**, *8*, 3947–3973. [CrossRef]
7. Krikken, F.; Schmeits, M.; Vlot, W.; Guemas, V.; Hazeleger, W. Skill improvement of dynamical seasonal Arctic sea ice forecasts. *Geophys. Res. Lett.* **2016**, *43*, 5124–5132. [CrossRef]
8. Troccoli, A. Seasonal climate forecasting. *Meteorol. Appl.* **2010**, *17*, 251–268. [CrossRef]
9. Peng, P.; Kumar, A.; Barnston, A.G.; Goddard, L. Simulation Skills of the SST-Forced Global Climate Variability of the NCEP–MRF9 and the Scripps–MPI ECHAM3 Models. *J. Clim.* **2000**, *13*, 3657–3679. [CrossRef]
10. van Oldenborgh, G.J.; Balmaseda, M.A.; Ferranti, L.; Stockdale, T.N.; Anderson, D.L.T. Did the ECMWF Seasonal Forecast Model Outperform Statistical ENSO Forecast Models over the Last 15 Years? *J. Clim.* **2005**, *18*, 3240–3249. [CrossRef]
11. Schepen, A.; Wang, Q.J.; Robertson, D.E. Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall. *J. Geophys. Res. Atmos.* **2012**, *117*, D20107. [CrossRef]
12. Zhang, W.; Villarini, G.; Vecchi, G.A.; Murakami, H.; Gudgel, R. Statistical-dynamical seasonal forecast of western North Pacific and East Asia landfalling tropical cyclones using the high-resolution GFDL FLOR coupled model. *J. Adv. Model. Earth Syst.* **2016**, *8*, 538–565. [CrossRef]
13. Folland, C.K.; Scaife, A.A.; Lindesay, J.; Stephenson, D.B. How potentially predictable is northern European winter climate a season ahead? *Int. J. Climatol.* **2012**, *32*, 801–818. [CrossRef]
14. Herman, G.R.; Schumacher, R.S. Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. *Mon. Weather Rev.* **2018**, *146*, 1571–1600. [CrossRef]

15. Weyn, J.A.; Durran, D.R.; Caruana, R. Can Machines Learn to Predict Weather? Using Deep Learning to Predict Gridded 500-hPa Geopotential Height From Historical Weather Data. *J. Adv. Model. Earth Syst.* **2019**, *11*, 2680–2693. [CrossRef]
16. Qian, S.; Chen, J.; Li, X.; Xu, C.Y.; Guo, S.; Chen, H.; Wu, X. Seasonal rainfall forecasting for the Yangtze River basin using statistical and dynamical models. *Int. J. Climatol.* **2020**, *40*, 361–377. [CrossRef]
17. Nooteboom, P.D.; Feng, Q.Y.; López, C.; Hernández-García, E.; Dijkstra, H.A. Using network theory and machine learning to predict El Niño. *Earth Syst. Dyn.* **2018**, *9*, 969–983. [CrossRef]
18. Nielsen, K.; Estella Perez, V.; Troccoli, A.; Calcagni, E.; Catalano, F.; Formento, M.; Geertsema, G.; Krikken, F.; Maksimovich, E.; Morgani, M.; et al. SECLI-FIRM Report D2.2: Report on the Role of Large-Scale Climate Phenomena and Teleconnections on the Predictability of the Key Predictands for the Case Study Applications (v2). Improving the Skill of Seasonal Forecasts through Multi-Model Combination, Advanced Statistical Methods, and Signal Boosting. 2021. Available online: https://www.secli-firm.eu/project-reports (accessed on 5 February 2024).
19. Fan, Y.; van den Dool, H. A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res. Atmos.* **2008**, *113*, D01103. [CrossRef]
20. Huang, B.; Thorne, P.W.; Banzon, V.F.; Boyer, T.; Chepurin, G.; Lawrimore, J.H.; Menne, M.J.; Smith, T.M.; Vose, R.S.; Zhang, H.M. Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons. *J. Clim.* **2017**, *30*, 8179–8205. [CrossRef]
21. Schneider, U.; Becker, A.; Finger, P.; Meyer-Christoffer, A.; Ziese, M.; Rudolf, B. GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. *Theor. Appl. Climatol.* **2014**, *115*, 15–40. [CrossRef]
22. Yin, Y.; Alves, O.; Oke, P.R. An Ensemble Ocean Data Assimilation System for Seasonal Prediction. *Mon. Weather Rev.* **2011**, *139*, 786–808. [CrossRef]
23. van Oldenborgh, G.J.; te Raa, L.A.; Dijkstra, H.A.; Philip, S.Y. Frequency- or amplitude-dependent effects of the Atlantic meridional overturning on the tropical Pacific Ocean. *Ocean Sci.* **2009**, *5*, 293–301. [CrossRef]
24. Meinshausen, M.; Smith, S.J.; Calvin, K.; Daniel, J.S.; Kainuma, M.L.T.; Lamarque, J.F.; Matsumoto, K.; Montzka, S.A.; Raper, S.C.B.; Riahi, K.; et al. The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Clim. Chang.* **2011**, *109*, 213. [CrossRef]
25. Mueller, B.; Seneviratne, S.I. Hot days induced by precipitation deficits at the global scale. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 12398–12403. [CrossRef] [PubMed]
26. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
27. van der Ploeg, T.; Austin, P.C.; Steyerberg, E.W. Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Med Res. Methodol.* **2014**, *14*, 137. [CrossRef] [PubMed]
28. (C3S), C.C.C.S. ERA5: Fifth Generation of ECMWF Atmospheric Reanalyses of the Global 355 Climate. Copernicus Climate Change Service Climate Data Store (CDS). 2017. Available online: https://cds.climate.copernicus.eu/ (accessed on 15 January 2021).
29. Kobayashi, S.; Ota, Y.; Harada, Y.; Ebita, A.; Moriya, M.; Onoda, H.; Onogi, K.; Kamahori, H.; Kobayashi, C.; Endo, H.; et al. The JRA-55 Reanalysis: General Specifications and Basic Characteristics. *J. Meteorol. Soc. Japan. Ser. II* **2015**, *93*, 5–48. [CrossRef]
30. Gelaro, R.; McCarty, W.; Suárez, M.J.; Todling, R.; Molod, A.; Takacs, L.; Randles, C.A.; Darmenov, A.; Bosilovich, M.G.; Reichle, R.; et al. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Clim.* **2017**, *30*, 5419–5454. [CrossRef] [PubMed]
31. Harris, I.; Jones, P.; Osborn, T.; Lister, D. Updated high-resolution grids of monthly climatic observations—The CRU TS3.10 Dataset. *Int. J. Climatol.* **2014**, *34*, 623–642. [CrossRef]
32. Lenssen, N.J.L.; Schmidt, G.A.; Hansen, J.E.; Menne, M.J.; Persin, A.; Ruedy, R.; Zyss, D. Improvements in the GISTEMP Uncertainty Model. *J. Geophys. Res. Atmos.* **2019**, *124*, 6307–6326. [CrossRef]
33. Becker, A.; Finger, P.; Meyer-Christoffer, A.; Rudolf, B.; Schamm, K.; Schneider, U.; Ziese, M. A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present. *Earth Syst. Sci. Data* **2013**, *5*, 71–99. [CrossRef]
34. Hagedorn, R.; Doblas-Reyes, F.J.; Palmer, T.N. The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus A Dyn. Meteorol. Oceanogr.* **2005**, *57*, 219–233. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.