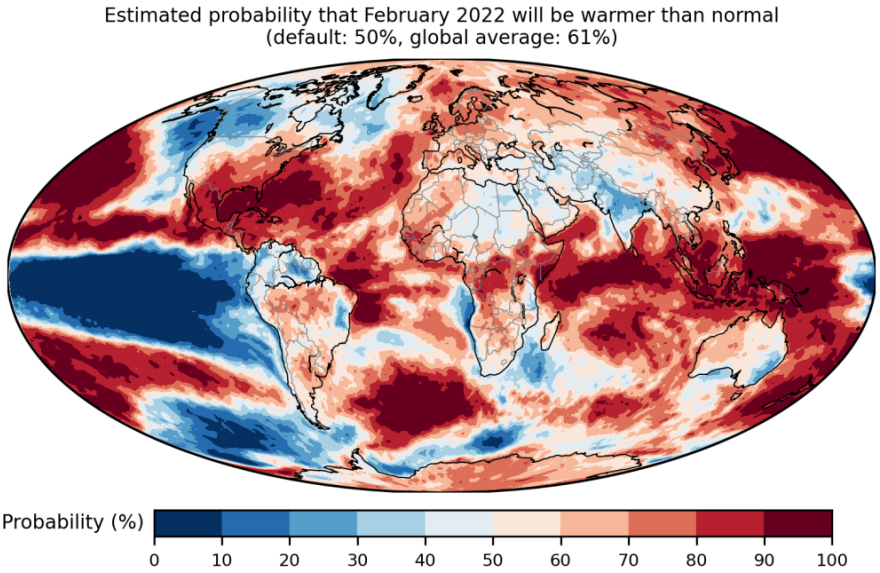
Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

A Benchmarking Dataset for Seasonal Weather Forecasts



Note no SAMBA/01/22

Authors Alex Lenkoski (NR)
Erik Kolstad (NORCE)
Thordis Thorarinsdottir (NR)

Date 27th January 2022

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**A Benchmarking Dataset for Seasonal Weather Forecasts**
Authors	**Alex Lenkoski (NR)** <alex@nr.no> **Erik Kolstad (NORCE)** <ekol@norceresearch.no> **Thordis Thorarinsdottir (NR)** <thordis@nr.no>
Date	27th January 2022
Publication number	SAMBA/01/22

Abstract

There is an increasing demand for high-quality seasonal weather forecasts from a broad range of stakeholders. However, the numerical weather prediction (NWP) output on which these forecasts are based require substantial postprocessing, as they are subject to systematic errors in both mean and spread. In order to validate any proposed post-processing methodology, the research community would benefit from a benchmark dataset on which more sophisticated methods can quickly be developed and tested. We supply a multi-model, multi-variable global dataset using five forecasting systems from the Copernicus climate data store (CDS) which can help serve these purposes. Our dataset is constructed using a straightforward anomaly standardization methodology with a leave-year-out cross validation design. In addition, validating observations from the ERA5 dataset are supplied, enabling rapid verification of system performance. The goal of this dataset is to save the research community the substantial investment in time necessary to create a usable baseline for their own investigations and also to create a standard benchmark dataset to which different research groups can compare results.

Keywords	Seasonal weather forecasting
Target group	General Public
Availability	
Project	
Project number	
Research field	
Number of pages	10

Contents

1 Introduction

Substantial effort has recently been placed on developing seasonal weather forecasting systems by the meteorological community (Johnson et al., 2019; MacLachlan et al., 2015). Given the ambitious scope of forecasting with accuracy at such long lead-times, it is clear that there will have to be concerted coordination between the numerical modeling and statistical post-processing communities. To aid in this development, the Copernicus Climate Data Store (CDS) allows meteorological organizations to upload the output of their seasonal forecasting systems to a centralized location and in a roughly consistent format. In addition, hindcasts are supplied, which allow for the correction of systematic model error (i.e. bias, variance and quantile correction). When assessing statistical post-processing techniques, these hindcasts also form the basis for out of sample performance validation.

Individual users of CDS can then download this output and use it to research their own seasonal forecasting methodology, as well as issue their own seasonal forecasts. The data hosted on the CDS are immense, and there can understandably be a delay in retrieving all hindcasts for a given meteorological quantity of interest from CDS. Furthermore, the raw format and occasional quirks of much of this output can require substantial additional effort on the part of a researcher in order to organize the data store for research purposes. Our goal has been to provide a single centralized dataset on which the majority of these steps have already been performed for monthly-mean forecasts for three important meteorological quantities, namely two meter temperature, total precipitation and ten meter wind speed.

Given the existence of multiple NWP seasonal systems, it is natural for researchers to consider multi-model post-processing methodologies. Furthermore, on the seasonal timescale, it is important to allow researchers to ask joint questions across multiple meteorological quantities, spatial areas and/or multiple time horizons. Ideally, these topics can be addressed in a probabilistic manner. Therefore, postprocessing methodologies must in some manner either retain the dependence structure present in the underlying raw model output, or layer a model for joint distributions on top of a univariate pooling methodology. In order to be useful, a benchmark dataset must therefore allow for both approaches to be entertained.

Seasonal NWP output is subject to persistent bias and variance issues that render model output essentially unusable if not corrected for, which obviates the need for statistical post processing. A host of post processing methodologies for seasonal NWP output have been proposed, see e.g. Hemri et al. (2020) for a comprehensive review of different techniques. Our objective with the proposed dataset is to perform a reasonable baseline post-processing on which more sophisticated methods can be developed. This approach not only renders the model output useful for downstream modeling, but also can serve as a benchmark post-processing methodology in its own right with which more sophisticated approaches can be compared quickly.

We use a straightforward mean and variance correction system, as the approach is parsimonious and yet still serves as a realistic baseline methodology. In particular, it retains

much of the multivariate structure of the ensemble system, which regression based approaches such as MOS would destroy. Furthermore in the context of surface temperature, Hemri et al. (2020) shows conclusively that local EMOS type methods performed poorly, partially due to a lack of substantial training data. The fact that our methodologies retain the key dependencies in the underlying ensemble is also useful in that it e.g. allows for joint analyses to be conducted without the need to elicit context-specific multivariate MOS distributions, or conduct the explicit joint postprocessing techniques outlined in Hemri et al. (2020).

The structure of this report is as follows. In Section 2 we discuss post processing methodologies for monthly mean values of seasonal forecasting systems. In Section 3 we discuss the data that are retrieved from CDS and subsequently post-processed. Section 4 then outlines the structure of the dataset. Section 5 concludes.

2 Methods

In this section we outline the statistical postprocessing of seasonal NWP output for monthly mean values. Our methodology is based on anomaly standardization. This reduces the output of an NWP monthly mean to a "standardized anomaly", which can then be rescaled and mean-adjusted for any chosen observational/reanalysis reference dataset. In our examples, we chose the ERA5 data as the relevant reference data.

Let $\mathcal{M}$ be a collection of NWP models which issue forecasts over a set of locations $\mathcal{S}$ for a collection of weather quantities $\mathcal{X}$. In the course of a year, each system has a collection of initialization times $\mathcal{T}$ and J_M ensemble members, which vary according to the model $M \in \mathcal{M}$ and a collection of years $\mathcal{Y}$ over which the model has been run. We focus on monthly mean values and each system supplies forecasts for L months ahead for each initialization date (In our case $L = 5$). Thus x_{msytlj} is the jth ensemble member of model m for weather quantity $x \in \mathcal{X}$ at location $s \in \mathcal{S}$ for initialization time $t \in \mathcal{T}$ and lead-month $l = \{0, \ldots, L\}$.

It is well established that on a seasonal timescale the raw output x_{msytlj} is rarely useful on its own due to systematic model bias. Thus, we will always post-process this model output. From the output x_{msytlj} we construct several summary statistics which are each specific to the system, initialization and lead-time in question. These are

- Climatology

$$\bar{x}_{mstl} = \frac{1}{|\mathcal{Y}|J_m} \sum_{y \in \mathcal{Y}} \sum_{j=1}^{J_m} x_{msytlj}$$

- Variance

$$\sigma^2_{mstl} = \frac{1}{|\mathcal{Y}|J_m - 1} \sum_{y \in \mathcal{Y}} \sum_{j=1}^{J_m} (x_{msytlj} - \bar{x}_{mstl})^2$$

- Raw Anomaly

$$\tilde{x}_{msytlj} = x_{msytlj} - \bar{x}_{mstl}$$

- Standardized Anomaly

$$\hat{x}_{msytl} = \tilde{x}_{msytlj} / \sqrt{\sigma^2_{mstl}}.$$

The collection of standardized anomalies can then be used to issue a probabilistic forecast based on the multi-model system output. Furthermore, as opposed to regression-based MOS methods, the dependence structure in the raw ensemble output is carried over to the standardized anomalies. This means that probabilistic inquiries on two separate quantities in $\mathcal{X}$, separate lead times, or spatial areas in $\mathcal{S}$ can be addressed even after standardization.

The process above yields a $\sum J_M$ member ensemble of monthly mean forecasts. From this ensemble, any number of summary quantities can be derived such as the mean level, the deciles of the distribution for each combination of $sytl$, i.e. on a location, forecast year, forecast time and lead time basis, for each variable in question. By also retaining which system a given standardized anomaly came from, it is possible to assess the contribution of each NWP model to the overall performance of the multi-model system and address the effect of ensemble size on performance.

The anomaly standardization procedure allows the model output to then be used relative to any observational product of interest. One obvious choice to which the output can be compared is the ERA5 reanalysis data, which are also accessible via CDS and on the same grid. In this case, comparing standardized anomalies of the ERA5 data to the standardized anomalies of the NWP output is equivalent to doing so on the rescaled level. Therefore, to allow for rapid validation of any proposed method, we supply the ERA5 data in a leave-one-year-out standardized anomaly format.

In particular consider the reference observation Y_{sny} for location $s \in \mathcal{S}$ year $y \in \mathcal{Y}$ and month n. In a leave-one-year-out context, these data are not observed and thus must be excluded from calculation of climatologies and standard deviations. Thus, we define the leave-year-out climatology and standard deviation as

$$\bar{Y}_{sny} = \frac{1}{|\mathcal{Y}| - 1} \sum_{w \in \mathcal{Y}: w \neq y} Y_{swn}$$
$$\sigma^2_{sny} = \frac{1}{|\mathcal{Y}| - 2} \sum_{w \in \mathcal{Y}: w \neq y} (Y_{swn} - \bar{Y}_{snwy})^2$$

We then supply the leave-year-out standardized anomaly as

$$\tilde{Y}_{sny} = (Y_{sny} - \bar{Y}_{sny}) / \sqrt{\sigma^2_{sny}}.$$

This enables various methodologies to be quickly compared relative to an established reanalysis dataset. In particular, the first "forecast" to which any methodology is compared is often the climatology. We note that in our framework, the climatology forecast is equivalent to a forecast of 0, and thus $\tilde{Y}_{sny}$ is also the leave-year-out error of the climatology forecast, which enables skill scores to be rapidly calculated.

3 Input Data and Derived Products

Our dataset uses the output from five "core" systems hosted on the CDS, namely the ECMWF, UKMO, Meteo France, CMCC and DWD forecasting systems. These systems report their raw, subdaily output on slightly different grids, but CDS natively converts all output for the globe to the 1° regular grid for monthly mean values. We use these converted datasets as our basis for development. We use the full global data, thus 65160 grid points and include month lead times 1 through 3 in our dataset, as these lead times are often of primary interest[1].

CDS Forecasts began being issued in January 2017, and for each forecast month hindcasts for the period 1993-2016 are also supplied. Typically, hindcasts for a given initialization month are first published alongside (or one month before) the associated forecast. In addition, ensemble sizes for the issued forecasts are typically larger than those supplied in the hindcast period. Furthermore, models are under constant development and thus each system is often updated several times (with only the hindcast period reissued). The dynamic nature of this data store is simultaneously realistic (in that it represents the reality of working with constantly evolving forecast products) and problematic from the perspective of conducting research on a consistent dataset.

We therefore only consider the hindcast period 1993-2016, which leads to a consistent ensemble size[2]. Furthermore, for a particular initialization month, we use the model number that was in use for that month in 2021.[3] Thus for the systems UKMO and Meteo France, this implies that different initialization months may have different system numbers associated with them. However, for all years in the period 1993-2016 the same set of model numbers are used for each initialization month. We feel this gives the benchmark dataset a desirable degree of consistency, while also representing the dynamic nature of the underlying forecasting framework that we are researching, in that periodic model changes are part of the forecasting environment.

In addition to these forecast data, we have collected the ERA5 reanalysis data on the same grid and for the associated period. Both the systems and reanalysis data are then converted into standardized anomalies as discussed in Section 2. We consider three surface variables in our dataset, namely two meter temperature, total precipitation and ten meter wind speed.

1. All data are downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store
2. The model output for January 1993 is missing for the UKMO model and is supplied as missing values in our dataset
3. In particular, for ECMWF this is system 5 throughout. For UKMO this is system 15 for January and February and system 600 for the remaining months. For Meteo France this is system 7 for January through June and system 8 for July through December. For DWD and CMCC this is system 21 and 35, respectively, for all months

4 Dataset structure

Our dataset is a single netcdf file containing six arrays[4]. Each array is named to indicate which meteorological variable it is associated with (`2m_temperature`, `total_precipitation`, `10m_windspeed`) and whether it is model output (`nwp`) or reanalysis data (`era`). Thus, the array `2m_temperature_nwp` contains the standardized anomalies for the NWP output for two meter temperature and `10m_windspeed_era` contains the array of leave year out standardized anomalies for the ERA5 data.

The `nwp` arrays all have six dimensions, namely `lon`, `lat`, `forecast_year`, `forecast_month`, `lead_month`, `global_ensemble_number`. An auxiliary variable, `system`, maps the global ensemble number to each of the five systems. The `era` arrays have five dimensions, which are the same as the `nwp` arrays, but excluding the global ensemble number dimension. We have structured the `era` arrays to align directly with the structure of the `nwp` arrays. This means that a given month of reanalysis data is repeated three times in this array. This is done to reduce what can otherwise be a rather tedious bookkeeping exercise to align a forecast year, forecast month, lead time triple with the associated observation year and month pair. This does imply that the `era` arrays are three times as large as is strictly necessary. However, as the netcdf file is compressed, the actual increase in size is substantially less and the convenience of this structure makes this, in our opinion, worth the extra size.

5 Conclusions

We have outlined a dataset whose primary purpose is to accelerate research into the statistical postprocessing of seasonal weather forecasts. The motivation for this dataset was our experience working with the CDS data store in its raw format. A substantial amount of time was spent simply organizing and processing these data, leaving less room for the investigation of methodologies.

Design choices for this dataset were therefore made primarily with speed in mind. It should now be possible for other researchers to almost immediately begin investigating new methodologies and ascertain the cross validated skill of these methods, both relative to the basic climatology forecast (which is conveniently 0 in our data for all quantities) and relative to a straightforward anomaly standardization approach. While considerable effort was spent to structure the dataset to allow for rapid investigations, we otherwise wanted to give the researcher as much flexibility as possible. We therefore chose to have a fully global dataset, to include a number of lead times and three important variables. Most importantly, we left the ensemble nature of the underlying systems intact, enabling a variety of analyses to be entertained.

At this stage, it should be relatively straightforward to perform a large number of inter-

4. Please note that this report is currently under submission at a journal specializing in the dissemination of scientific datasets. The intention is for the dataset to then be published and hosted via this journal. Individuals interested in receiving the dataset before this process has been completed are welcome to contact the first author directly

esting analyses. Since the system to which each global ensemble member is associated has been retained, a natural investigation would be the relative contribution to skill of each forecasting system. Questions related to the forecasting skill of compound events across different meteorological quantities, geographic areas or lead times can also be considered. Furthermore, these data would be amenable to the investigation of more involved machine learning approaches to post processing.

References

DelSole, T., Yang, X., and Tippett, M. K. (2013). Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quarterly Journal of the Royal Meteorological Society*, 139(670):176–183.

Hemri, S., Bhend, J., Liniger, M. A., Manzanas, R., Siegert, S., Stephenson, D. B., Gutiérrez, J. M., Brookshaw, A., and Doblas-Reyes, F. J. (2020). How to create an operational multi-model of seasonal forecasts? *Climate Dynamics*, 55(5):1141–1157. 5, 6

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., et al. (2019). Seas5: the new ecmwf seasonal forecast system. *Geoscientific Model Development*, 12(3):1087–1117. 5

MacLachlan, C., Arribas, A., Peterson, K. A., Maidens, A., Fereday, D., Scaife, A., Gordon, M., Vellinga, M., Williams, A., Comer, R., et al. (2015). Global seasonal forecast system version 5 (glosea5): A high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, 141(689):1072–1084. 5