Environmental Modelling & Software 103 (2018) 29–42

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

An R package for climate forecast verification

Nicolau Manubens [a, *], Louis-Philippe Caron [a], Alasdair Hunter [a], Omar Bellprat [a], Eleftheria Exarchou [a], Neven S. Fučkar [a, e], Javier Garcia-Serrano [a], François Massonnet [a, b], Martin Ménégoz [a], Valentina Sicardi [a], Lauriane Batté [c], Chloé Prodhomme [a], Verónica Torralba [a], Nicola Cortesi [a], Oriol Mula-Valls [a], Kim Serradell [a], Virginie Guemas [a, c], Francisco J. Doblas-Reyes [a, d]

[a] Barcelona Supercomputing Center (BSC), Barcelona, Spain
[b] Université Catholique de Louvain, Louvain-la-Neuve, Belgium
[c] CNRM UMR 3589, Météo-France/CNRS, Toulouse, France
[d] Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010, Barcelona, Spain
[e] Environmental Change Institute, University of Oxford, Oxford, UK

ARTICLE INFO

Article history:
Received 22 February 2017
Received in revised form
12 January 2018
Accepted 17 January 2018
Available online 22 February 2018

Keywords:
Forecast verification
Climate forecast
Skill score
R package

ABSTRACT

Forecast verification is necessary to determine the skill and quality of a forecasting system, and whether it shows improvement with pre- or post-processing. *s2dverification* v2.8.0 is an open-source R package for the quality assessment of climate forecasts using state-of-the-art verification scores. The package provides tools for each step of the forecast verification process: data retrieval, processing, calculation of verification measures and visualisation of the results. Examples are provided and explained for each of these stages using climate model output.

1. Introduction

From their modest beginning in the 1950s, numerical weather forecasts have evolved in scale and complexity to become an integral part of countless decision making processes worldwide. Although not as widely disseminated as weather forecasts, many operational research centres also produce seasonal climate forecasts (Doblas-Reyes et al., 2013). These forecasts estimate the monthly and seasonal mean values of climate variables (e.g. temperature and precipitation) 1–12 months in advance to anticipate, for example, drought or flooding events related to large scale atmospheric flow patterns, such as the El Niño Southern Oscillation (ENSO; Rasmusson and Wallace (1983); Kousky et al. (1984)) or the North Atlantic Oscillation (NAO; Hurrell (1996)). With increases in computing power combined with advances in climate science and in the quality and quantity of observational data, there is growing interest in the feasibility of decadal climate forecasts (commonly referred to as climate forecasts), i.e. forecasts providing estimates one year to several decades in advance (Smith et al., 2007). The accuracy of these forecasts has progressed substantially over the last 20 years (e.g. see Doblas-Reyes et al. (2013) and Meehl et al. (2014), for more information on seasonal and decadal forecasts, respectively).

Regardless of the timescale in consideration, it is essential to assess the forecast's quality through forecast verification. Forecast verification is conducted by comparing forecasts of past events (also known as retrospective forecasts or hindcasts) with their corresponding observations (or observational products; referred to as references hereafter). The verification involves quantifying the accuracy (the correspondence between the forecasts and references) and the association (the strength of the relationship between the forecasts and the references) (Potts, 2003) of the forecast system. Ideally a number of different metrics should be considered, as a single measure cannot fully characterise the forecast quality (Bennett et al., 2013).

This manuscript introduces *s2dverification* v2.8.0 (where s2d

* Corresponding author.
E-mail address: nicolau.manubens@bsc.es (N. Manubens).

https://doi.org/10.1016/j.envsoft.2018.01.018
1364-8152/

stands for seasonal to decadal), a software package for the *R* statistical programming language (R Core Team, 2015). The key advantage of using *s2dverification* is that it enables the user to conduct the entire workflow of the verification process with a single software package. It provides an end-to-end unified framework, including a powerful data loading and homogenization function (Section 2.1), functions that transparently implement well known methods for pre-processing and verifying climate data, novel verification and synthetic forecast generation methods, and functions to generate frequently required visual products. The package stems from a collection of verification tools developed over several years by scientists from the Earth Sciences Department of the Barcelona Supercomputing Center-Centro Nacional de Supercomputación and their collaborators. The group specializes in the production and analysis of climate forecasts derived from dynamical climate models or statistical forecast systems with forecast periods ranging from weeks to years. The package has also been developed in conjunction with external partners to address the needs of forecast users from the private sector. Although the package is designed mainly for the verification of forecasts of any climate variable, it can also be used for forecast verification in other fields or on different timescales. The groups of functions and modules in *s2dverification* and their potential interactions are depicted in Fig. 1.

Several software packages exist for weather and climate forecast verification, for a variety of platforms and programming languages. These range greatly in the breadth of their contents, from solely calculating verification statistics to providing a complete framework including tools for data management, visualisation and other analyses typical of the field the package specializes in. Outside of R, some of the most relevant open source packages that provide a full framework for the evaluation of climate model output are the Model Evaluation Tools (MET; Brown et al. (2009)) developed at the National Center for Atmospheric Research, U.S. (NCAR), the Ensemble Verification System (EVS; Brown et al. (2010)) developed at the National Oceanic and Atmospheric Administration, U.S. (NOAA) and the Earth System Model Validation Tool (ESMValTool; Eyring et al. (2015)), which provides community-reviewed tools for calculating metrics for simulations produced in the context of the Coupled Model Intercomparison Project (CMIP).

A summary of some of the available R packages containing functions relevant to forecast verification is provided in Table 1. The

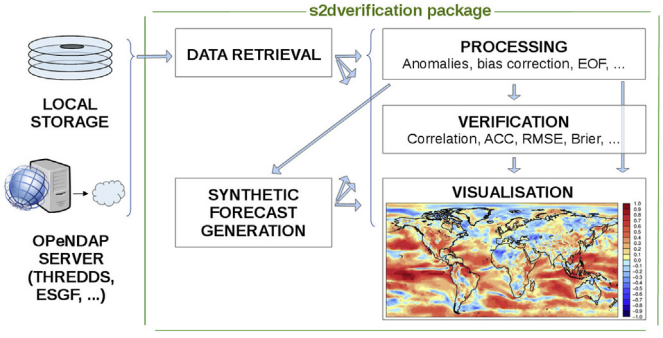

Fig. 1. The core modules of the *s2dverification* software package. The data retrieval module gathers the forecast and reference data required for the analysis from either local storage or remote OPeNDAP (https://www.opendap.org/) compliant servers (see Section 2.1), which can then be passed directly to any of the other modules; data can also be synthetically generated by the synthetic forecast generation module (see Section 2.5); the processing module pre-processes or performs preliminary analyses of the data, such as computing the drift-corrected predictions or modes of variability; the verification module computes deterministic or probabilistic scores or skill scores; data from any of these modules can be plotted with the visualisation module, which generates plots of either time series, maps or map animations.

added-value of *s2dverification* is its data management utilities, designed for the efficient computation of forecast quality, its range of state-of-the-art forecast quality evaluation scores, and visualisation tools tailored to forecast quality inspection.

The rest of this paper is organized as follows. Section 2 describes the main modules the package comprises; data retrieval, processing, verification measures and visualization. Functions for generating synthetic data and verifying hurricane forecasts are also described in this section. Section 3 illustrates the use of the package with a case study utilizing climate model output. Section 4 concludes this paper, and discusses some future developments for the package.

2. Methods

The *s2dverification* package provides tools to assist with each of the four stages of the verification process; data retrieval, processing, calculation of verification measures and visualization. The package is constantly evolving, with new verification measures and functionality being added. In this section we present examples of some of the available functions for each stage of the verification process. Some of the core *s2dverification* functions are listed in Table 2.

2.1. Data retrieval

Data retrieval refers to the loading and homogenizing of the data sets so that they are in a commensurable format which is compatible with the processing and analysis functions. The data retrieval stage can be relatively time consuming due to the following reasons: i) it may involve several software packages, ii) the data sets may come from different sources and follow different conventions, and finally iii) data may need to be interpolated onto a common grid. The *Load* function in *s2dverification* streamlines this process, automating many of the steps, provided that the data are stored according to some simple guidelines. These guidelines are as follows:

- The data is in NetCDF format in a local file system or on remote servers that are able to communicate using the Open-source Project for a Network Data Access Protocol (OPeNDAP), such as THematic Real-time Environmental Distributed Data Services[1] (THREDDS) servers or Earth System Grid Federation[2] (ESGF) nodes.
- There is one file for each simulation initialized at a different time (referred to here as start date), with an identifier for the given start date appearing in the file path or filename.
- The reference data contains either one file per month, with the corresponding year and month included in its path or filename, or a single file for the entire reference data set.
- All files under consideration contain daily or monthly area-averages or two-dimensional fields on either regular rectangular or Gaussian grids.

The *Load* function can be used to retrieve data from various forecast systems and reference data sets for multiple start dates, model members, and forecast time steps, for a user-defined region and variable. The user can interpolate the data on a common grid, choosing between the interpolation techniques supported by the Climate Data Operators software (CDO; Schulzweida (2015)). The *Load* function also allows the user to apply separate masks to both

[1] http://www.unidata.ucar.edu/software/thredds/current/tds/.
[2] http://esgf.llnl.gov/.

Table 1
Overview of R packages currently available for assisting with the processing, verification and visualization of climate model output. References for relevant papers have been provided, when available. For all other packages please visit the CRAN homepage https://cran.r-project.org/.

package	Functionality	Reference
raster, sp, zoo	Handling and visualization of spatio-temporal data sets.	Hijmans et al. (2015)
wux, loadeR	Collection and homogenization of data sets for the verification process.	Mendlik et al. (2015)
remote, SpatialVx	Analysis, verification and visualization of spatial data.	Appelhans et al. (2015), Gilleland and Gilleland (2016)
SpecsVerification, easyVerification	Verification routines for deterministic and probabilistic forecasts.	Siegert (2015), Bhend (2016).
esd, downscaleR	Frameworks for the empirical-statistical downscaling of weather and climate variables, tools for climate data analysis and visualisation.	Benestad et al. (2016), Group (2015)

Table 2
Main functions in the *s2dverification* package.

Function	Module	Description
Load	Retrieval	Function for loading, pairing and homogenizing forecast and reference data sets into R.
Clim	Processing	Calculates the climatologies of the data sets output by Load.
EOF	Processing	Calculates the area-weighted empirical orthogonal functions (EOF) using singular value decomposition (SVD).
Corr, RMS, RatioRMS	Verification	Calculate the correlation, root-mean-square-error (RMSE), and ratio of the RMSEs, respectively.
UltimateBrier	Verification	Function for calculating the Brier score and its decomposition for a multi-model forecast and references.
PlotEquiMap	Visualization	Function for plotting maps of variables or verification statistics.
PlotAno	Visualization	Function for plotting time series of anomalies.

the forecasts and references, either when requesting two-dimensional fields or before averaging data over a specific region. The data are loaded into two multi-dimensional arrays of similar structure: one containing the forecasts and the other containing the references. The data pairing, i.e. searching for the references matching the dates of the forecasts, is done automatically. The resulting arrays are the working units of *s2dverification* that can then be passed directly to other functions in the package.

This powerful feature significantly reduces the amount of time spent manipulating files since any modifications to the desired output (e.g. the inclusion of a new reference data set) are handled by *Load* and it is straightforward to modify the parameters in the function call.

Load is designed to work on a single workstation connected to the file systems or servers containing the relevant data. Usually, in an infrastructure with this topology, the connection between the workstation and the data servers becomes a bottleneck, and it is hence recommended to minimize the amount of time the connection is unused. However, by default, the *Load* function is alternating data retrieval processes (where the connection is in use) with corresponding interpolation and arrangement processes (where the connection is unused), in a way that the connection is not optimally used. In order to address this, the function can be adjusted, via a parameter, to work on multiple cores. This way, when the connection to the sources of data is slow, *Load* runs multiple retrieval-processing couples simultaneously to constantly exploit the connection bandwidth. When the connection is fast, running *Load* in parallel will ensure the available computing resources in the workstation are used efficiently to interpolate and arrange the large flow of retrieved data as it is loaded.

The use of the CDO libraries in *s2dverification* makes it a useful and unique package for handling data on regular Gaussian grids, which are widely used in climate modelling, and for interpolating to or from regular longitude-latitude grids with a variety of methods.

2.2. Processing

One of the purposes of *s2dverification* is to provide a versatile post-processing adjustment of seasonal to decadal climate predictions through the application of different bias correction methods. Due to inherent model approximations, as well as errors in the initial conditions, boundary conditions and forcing fields, after initialization a climate model drifts towards its preferred climate state or attractor (Dijkstra, 2013). The model bias is an average model error over the validation period. The bias in climate forecasts can change in time exhibiting non-stationary drifts that can be dependent on the forecast time and start dates. The *Clim* function allows the user to account for the drift when calculating the climatology with one of three available empirical adjustment methods; i) the mean, per-pair bias correction method (e.g. García-Serrano and Doblas-Reyes (2012)); ii) the linear trend bias correction method Kharin et al. (2012); and iii) the initial condition bias correction method (Fučkar et al., 2014).

As well as bias correction, filtering and smoothing of the data is often necessary. The variability in a climate field is exhibited by a number of processes occurring at different time scales. For example, the persistence of an atmospheric anomaly is not longer than a few days while the persistence of a sea surface temperature anomaly can last for months or seasons. Internal (e.g. deep oceanic circulation) and external (e.g. solar activity or greenhouse gases) agents can drive variability at different frequencies via dynamical mechanisms. Depending on the type of analysis being conducted, some of these processes may be of greater interest than others, and it may be necessary to filter out some of that variability before analysing the data. Several procedures can be employed for this purpose:

- The Fourier transform filter, which is based on a discrete Fourier transform, and converts a finite series of samples equally spaced in time from its original time dimension into the frequency domain in order to ease the isolation of a specified frequency (e.g. Von Storch and Zwiers (2001)). *s2dverification* provides the functions *Spectrum* and *Filter*, which convert data to the frequency domain and filter out specified frequencies, respectively.
- The moving average filter (also known as the running mean), which computes the mean within a sliding time window, thereby representing a low-pass filter; e.g. in decadal prediction, a 4-year forecast moving average is often performed in order to retain interannual to decadal variability and to reduce

unpredictable higher frequency variability (García-Serrano and Doblas-Reyes, 2012). The package includes the *Smoothing* function to perform this operation on multi-dimensional arrays.
- The piecewise filter, which is applied to raw values and computes differences between consecutive samples, thus largely attenuating low-frequency signals; e.g. using yearly data it emphasizes inter annual anomalies (the departure from the long-term average, see e.g. Stephenson et al. (2000)), and using daily data it retains synoptic activity (24h-filter; e.g. Wallace et al. (1988)). For example, one could assess the characteristic transient activity of the extra-tropical storm tracks. Although there is no function in *s2dverification* for piecewise filtering, this simple procedure can be handled with the base R functions.

The forecast and reference data sets are typically very large, making it difficult to analyse individual processes and extract meaningful information from the data. However, a number of techniques are available for dimensionality reduction and data compression as well as pattern extraction to facilitate this purpose. Widely used methods include the calculation of the Empirical Orthogonal Functions (EOF), which identify the leading modes of variability in a particular field (e.g. Von Storch and Zwiers (2001)); and the calculation of Extended EOFs (Venegas, 2001), which apply EOFs to a time-evolving field. These can be computed with the *EOF* function in *s2dverification*. Climate data fields can then be regressed onto the leading modes of variability identified by *EOF* using the *ProjectField* function. The *SVD* function allows the user to investigate modes of covariability between two fields using Maximum Covariance Analysis (MCA) and Canonical Correlation Analysis (CCA) (Bretherton et al., 1992); as well as Extended MCA for the analysis of time-evolving coupled patterns (García-Serrano et al., 2008). The existence of groups, or clusters, within groups whose identities or patterns are not known in advance (e.g. Fučkar et al. (2016); Caron et al. (2015)), can be investigated with the *Clusters* function.

2.3. Verification

After processing, the use of the *s2dverification* package to calculate relevant statistics and scores to assist in the verification of the forecast is straightforward. All of the functions in the package are designed to take multidimensional arrays as input. The dimensions of the array can include multiple runs (ensemble members) from several climate models, as well as forecast time steps, initialization dates, geographical locations (longitudes and latitudes) and the scores can be computed along any of these dimensions. Usually in forecast verification the scores are computed along the start date dimension. Here we consider the verification of deterministic and probabilistic forecasts separately. In the context of ensemble forecasts, an example of a deterministic forecast is the ensemble mean (e.g. Nino 3.4 region sea-surface temperature (SST) will be 0.6 $^{\circ}C$ above the climatological average in winter) and an example of a probabilistic forecast could be the probability density function of the SST distribution, taking into account all ensemble members as possible outcomes.

2.3.1. Deterministic forecasts

The correlation coefficient is used to measure the linear dependence between deterministic forecasts and the corresponding references. It can be calculated for large arrays with the *Corr* function in *s2dverification*. In addition to Pearson's correlation, the non-parametric Spearman and Kendall rank correlation coefficients are also supported. As the correlation is a measure of association and not of accuracy, it is not sensitive to the forecast bias (the difference between the mean forecast and the mean reference) or to systematic errors. Consequently a forecast system providing anomalies of the same sign as the references will result in forecasts with very high correlations even if the amplitude of the anomalies is systematically lower or higher than the reference amplitude (Déqué, 2003).

The accuracy of a forecast system can be assessed with the root mean square error (RMSE). The *RMS* function calculates the RMSE between the forecast and reference anomalies. In addition to *RMS*, *RatioRMS* computes the ratio of RMSE scores of two different forecast data sets (see Appendix A). A ratio lower than one indicates that the forecast system used to generate the first set of forecasts performs better than the one used to generate the second set of forecasts. While the ratio between RMSEs is used to compare two forecast systems, the RMSE skill score (RMSSS), computed with the *RMSSS* function, can be used to assess the skill of the forecast with respect to a climatological forecast (a forecast based only on the climatological statistics). This score ranges from minus infinity (no skill) to 1 (perfect forecast). A score larger than 0 indicates an improvement with respect to the climatological forecast.

As well as being a measure of the forecast accuracy, the RMSE also provides information about the reliability of the ensemble spread. A forecast ensemble is considered reliable when it samples the full range of possible outcomes that can arise from a given initial state. If the forecast system has known imperfections, which are reflected in the RMSE, the ensemble spread should sample this range of model errors in order to represent the full range of possible outcomes. In practice, the ensemble spread is often smaller than the RMSE, and the forecasts are therefore overconfident. The diagnostic *RatioSDRMS*, known as the spread versus error relationship (Slingo and Palmer, 2011), measures this relationship by computing the fraction of the ensemble spread (standard deviation of the ensemble across all start dates) with the RMSE of the ensemble mean. A forecast is considered reliable if this measure is close to 1.

The ability of a forecast to reproduce spatial patterns can be investigated with the anomaly correlation coefficient, which is calculated with the *ACC* function (Krishnamurti et al., 2003). This function calculates the anomaly correlation coefficient to measure the linear association between the longitude-latitude forecast anomalies at a single time step and the corresponding observational reference anomalies. This is somewhat different from the function *Corr*, which evaluates the ability of a forecast system to capture temporal variability. Spatial correlation, like time correlation, is not sensitive to the forecast bias or to systematic errors in the forecast system variance.

2.3.2. Probabilistic forecasts

Probabilistic verification measures can be broadly divided into dichotomous (binary) or continuous measures. Dichotomous scores measure the accuracy of a forecast predicting the probability of a binary event, e.g. “Is it going to rain more than 300 mm this summer?”, whereas continuous scores measure the accuracy of the estimated probability at all possible thresholds, hence evaluating the entire continuous probability distribution function. In *s2dverification* the dichotomous Brier score can be calculated to measure the mean distance (for a given binary event) between the forecasts and references in probability space. Since the Brier score is dichotomous, it cannot be applied directly to a deterministic forecast. First, the probability of the deterministic ensemble forecast of exceeding a certain threshold needs to be computed. The function *ProbBins* allows the computation of such probabilities, which can then be provided as inputs to the function *BrierScore* to compute the Brier score. The function *UltimateBrier* allows one to use different versions of the Brier score that take into account corrections for limited ensemble size (Ferro, 2014) and

limited hindcast size (Ferro and Fricker, 2012), or to compute the decomposition of the Brier score as described in Stephenson et al. (2008).

2.3.3. Confidence intervals

When conducting hypothesis tests or calculating confidence intervals, serial dependence in the data needs to be taken into account because, for highly autocorrelated data, the number of independent measurements might be significantly lower than the actual number of data points (Trenberth, 1984). Failing to take this feature of the data into consideration will lead to a systematic overestimation of the significance of the skill scores.

In *s2dverification*, the *Eno* function computes the number of independent observations in a time series (also called effective sample size), and *EnoNew* performs the same computation but offers the option to filter out the trend or to exclude some frequency peaks before estimating the equivalent number of independent data (Guemas et al., 2014). *Eno* is transparently used in the calculation of the confidence intervals by the following functions: *ACC*, *Corr*, *RatioRMS*, *RatioSDRMS*, *RMS* and *RMSSS*. This approach allows the accurate computation of the statistical significance of skill scores when serial dependence is present in the data.

2.4. Visualization

Due to the large amount of data and variables that are usually involved in the verification process, visualizing the data in a comprehensible way is non-trivial. For example, about 10^9 values have to be stored and organized to encode the 2-m air temperature on a 512×256 global grid, for two forecast systems, each running with 10 members, initialized twice a year for 30 years and integrated over 6 months. Visualization tools are thus required in all of the stages of the verification process; to quickly inspect the results of a newly produced set of simulations (i.e. to check the physical consistency of the results), to assess the model output after processing, and to display verification statistics and confidence intervals in a user-friendly way.

s2dverification includes a set of tools to plot indices, scores and maps of user defined regions. Examples of usage and outputs of the most relevant plotting functions are shown in the next section. Time series plots for multiple models on the same axis can be created with the *PlotVsLTime* function, or of two different scores for multiple forecast systems against multiple experiments/observational references can be compared using the function *Plot2VarsVsLTime*. The function *PlotACC* plots the spatial anomaly correlation coefficients as a function of the start date and, if requested, also the forecast time, including the 95% confidence interval for the correlation in order to determine the statistical significance. *PlotClim* displays climatologies of a variable or index from multiple forecasts and references along the forecast time. *PlotAno* can display time series of either raw data or their deviation with respect to the climatology. For indices of teleconnection patterns, such as the NAO, information on the ensemble spread and correlation between the ensemble mean index and reference data index can be summarized visually using the *PlotBoxWhisker* function. This function plots for each start date the box-and-whisker diagram of the interquartile range, ensemble standard deviation and outlier members of the ensemble forecast, as well as the corresponding reference time series, based on outputs from the *EOF* function.

Spatial fields can be displayed on maps with cylindrical equidistant and stereographic projections with the *PlotEquiMap* and *PlotStereoMap* functions, respectively. The boundaries of the region to be plotted can be customized, the colourbar limits adjusted, continents shaded and arrows displayed, along with other customizable features. It is also possible to create animations with *AnimVsLTime*, a function that makes the composite of different maps in a GIF object in order to represent the time evolution of the provided feature, for instance the change in correlation indices with an observational reference as the forecast time increases. Finally, *PlotSection* plots a 2-dimensional subset (latitude-depth or longitude-depth) of a three-dimensional variable, for example the temperature in the atmosphere. The functions for plotting time series are based on R base's *plot*, *legend* and *boxplot*. The map plots consist of raster plots from R base's *image* function, superimposed to map projections drawn with the *maps* (Brownrigg, 2013) and *mapproj* (Deckmyn, 2015) or *GEOmap* and *geomapdata* (Lees, 2012) packages. The software ImageMagik is required as a system dependency for generating map animations in *AnimVsLTime*.

Several other packages have functions for plotting time-series, e.g. *ggplot2*, or *maps* with superimposed *raster* data, e.g. *ggplot2*, *lattice*, *sp*, *tmap* and *rasterVis*. The advantage of the visualisation functions provided in *s2dverification* are their ease of use, as they have been designed for application to data in the common *s2dverification* format or in raw R arrays.

2.5. Synthetic forecast generation

As well as functions for loading climate model output, *s2dverification* contains two functions for generating synthetic data from existing fields or by random sampling of variables from distributions defined by the user.

The *ToyModel* function is a statistical toy model based on the model presented in Weigel et al. (2008), with an extension to allow simulation from non-stationary distributions containing a linear trend. This toy model allows the exploration of forecast features for which little information is available, usually due to the reduced length of the hindcasts or ensemble members of real forecast systems. For example, how the correlation of the ensemble mean with the reference changes with increasing ensemble size (up to thousands of members). Further applications of this model can be found in Weigel et al. (2008), Siegert et al. (2015) and Bellprat and Doblas-Reyes (2016). The toy model imitates the typical components of a forecast: (i) predictability; (ii) forecast error; (iii) non-stationarity; and (iv) ensemble generation. The predictability is defined by the fraction of the observed outcome which is explained, hence the model does not serve as a probabilistic prediction but merely mimics different forecast components. It allows the generation of an artificial forecast based on some inputted references (obtained with *Load* or otherwise) or synthetic references based on the input parameters (standard deviation of interannual variability and a trend term). The model allows the user to vary the predictability of the system with a linear term from the reference anomalies and the magnitude of the forecast error. The magnitude of these terms is constrained by imposing the condition that the model has the same total variability as the references (see Weigel et al. (2008) for details). Imposing this condition allows the user to explore verification aspects which are free from systematic biases in the mean climate state and variability. The *Toymodel* function also includes parameters for defining the number of predictions, the ensemble size, the forecast length and the long-term trend.

2.6. Hurricane forecasting

The function *StatSeasAtlHurr* can be used to estimate the mean seasonal Atlantic hurricane activity in climate simulations. Hurricane activity is estimated using a statistical downscaling technique which relies on seasonal averages of the sea surface temperature

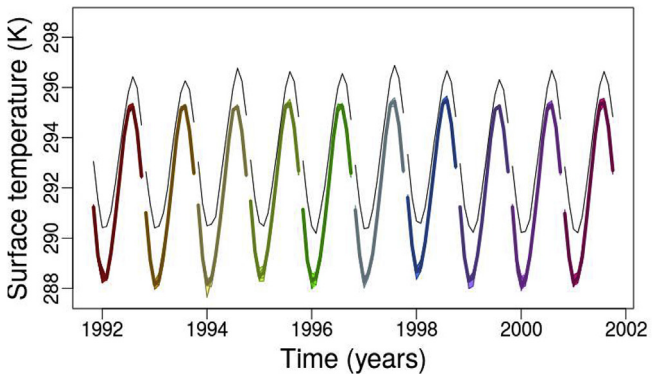

Fig. 2. Raw 2 m air temperature (tas, K) data from Experiment A. The black curve shows the corresponding references. Each start date is drawn in a different colour. Each member is drawn with a thin light line and the ensemble mean appears as a bold line in the same colour.

anomalies over the tropical Atlantic and over the tropics in general. There are dynamical and thermodynamical arguments linking sea surface temperature over these regions and Atlantic hurricane activity (Vecchi et al., 2011). More information on that link, on the downscaling technique itself and how that function can be used in hurricane forecast studies can be found in Villarini et al. (2010),Villarini and Vecchi (2012), Villarini et al. (2012) and Caron et al. (2014). *StatSeasAtlHurr* can be used to estimate either the mean number of hurricanes, the mean number of tropical cyclones with lifetime greater than 48 h or the mean power dissipation index (Emanuel, 2005), which is a measure which incorporates the number, intensity and duration of the storms over an entire hurricane season. The function also provides information on the distribution around the seasonal mean.

3. Case study

This section presents a case study which illustrates the use of some of the key functions for each of the four stages of the verification process. Forecasts from two different climate forecast systems are compared with an observational reference data set. In this example, near surface temperature (tas) forecasts have been selected, although the package would handle any other user-selected variable. Both forecast systems use the EC-Earth (v2.3.0) climate model, initialized every November 1st between 1991 and 2000. For the first forecast system, ocean initial conditions are taken from the ORAS4 reanalysis (Balmaseda et al., 2013) interpolated to the model grid, while for the second forecast system the ocean initial conditions are taken from an assimilation run which has been nudged to the ORAS4 reanalysis. Both forecast systems use atmosphere and land initial conditions from the ERA-interim reanalysis (Dee et al., 2011) and sea ice initial conditions from an ocean and sea ice coupled simulation forced by the Drakkar Forcing Set v4.3 (Brodeau et al., 2010). Five members are generated for each start date from perturbations of the initial ocean and atmosphere conditions (Du et al., 2012). The forecast duration is one year and the observational references are taken from the Drakkar Forcing Set v5.2 reanalysis (Dussin et al., 2014). This study considers the North Pacific region (10S-60°*E*, 100E-250°*E*).[3] The verification case study consists of the following steps.

- Loading and visualizing area averaged surface temperatures over the North Pacific from a set of forecasts and references.
- Computing and visualizing climatologies and anomalies.
- Computing and visualizing the correlation and RMSE between the forecasts and references.
- Loading two-dimensional subsets of the data over the North Pacific.
- Computing and visualising the modes of variability of the target data sets.
- Computing and visualising the spatial correlations and Brier scores.

3.1. Data retrieval

First, we construct lists containing the location of the data sets to be considered. The directory structure of the considered data files is represented in Appendix B. The path component must contain a character string with the path patterns to the files of the corresponding data set. The names surrounded by $ symbols are wildcards that *Load* will replace automatically with the appropriate value. Additional details can be found in *?Load*.

```
library(s2dverification)

expA <- list(name = 'experimentA',
             path = file.path('/path/to/experiments/$EXP_NAME$',
             'monthly_mean/$VAR_NAME$',
             '$VAR_NAME$_$START_DATE$.nc'))
expB <- list(name = 'experimentB',
             path = file.path('/path/to/experiments/$EXP_NAME$',
             'monthly_mean/$VAR_NAME$',
             '$VAR_NAME$_$START_DATE$.nc'))
obsX <- list(name = 'observationX',
             path = file.path('/path/to/observations/$OBS_NAME$',
             'monthly_mean/$VAR_NAME$',
             '$VAR_NAME$_$YEAR$$MONTH$.nc'))
```

[3] All the monthly mean files used in the example can be downloaded from www.bsc.es/projects/earthscience/s2dverification/s2dv_example_files.tar.

The *Load* command can now be called by specifying the variable of interest, the desired data sets, the region of interest, the start dates and the forecast time steps of interest. Here the area averages are requested with the *output* option:

```
sdates <- paste0(1991:2000, '1101')
data <- Load('tas', list(expA, expB), list(obsX),
             sdates = sdates,
             leadtimemin = 1, leadtimemax = 12,
             latmin = -10, latmax = 60,
             lonmin = 100, lonmax = 250,
             output = 'areave')
```

The object returned by *Load* contains two arrays; one for the forecast data (*data*$*mod*) and the other for the corresponding reference data (*data*$*obs*), with a dimension structure specific to *s2dverification*[4]:

```
c(n. of data sets, n. of members, n. of start dates,
n. of forecast time steps)
```

Before processing the data, we can first visually inspect the time series of the raw data using *PlotAno*:

```
PlotAno(Subset(data$mod, 'dataset', 1), data$obs, sdates,
        ytitle = 'surface temperature (K)',
        fileout = 'raw_expA_obsX.jpeg')
```

From Fig. 2, the forecast anomalies appear to be in broad agreement with the references, from which we can infer the data (both experimental and reference) have been correctly retrieved, and we can now proceed with the processing and verification.

3.2. Processing

Climatologies can be computed and plotted with *Clim* and *PlotClim*, respectively. *Clim* computes per-pair climatologies by default. Since by default *PlotClim* assumes that the forecasts start in January, the actual initial month in our experiments, November, needs to be specified with the parameter monini:

```
clim <- Clim(data$mod, data$obs)
PlotClim(clim$clim_exp, clim$clim_obs, monini = 11,
         ytitle = 'Surface temperature (K)',
         listexp = c('Experiment A', 'Experiment B'),
         listobs = c('Observation X'),
         fileout = 'clim_expA_expB_obsX.jpeg')
```

Each line in Fig. 3 shows the climatology of one member. A single climatology of the ensemble mean could be obtained by setting the parameter *memb* = *FALSE* in *Clim*. We can then obtain the anomalies with the function *Ano*, which subtracts the climatologies from each start date and member of the raw data. The anomalies can be plotted with *PlotAno* (see Fig. 4):

```
ano_mod <- Ano(data$mod, clim$clim_exp)
ano_obs <- Ano(data$obs, clim$clim_obs)
# The Subset function takes a slice of the data set. In this particular
# example, it is taking the anomaly data of the first experimental data set.
PlotAno(Subset(ano_mod, 'dataset', 1), ano_obs, sdates,
        ytitle = 'Surface temperature (K)', linezero = TRUE,
        fileout = 'ano_expA_obsX.jpeg')
```

By default, all the forecasts started at different dates contribute to the per-pair climatology and, when computing the anomaly of one of the start dates, its contribution to the climatology is subtracted. In order to avoid taking into account this contribution, thus mimicking as closely as possible an operational context in which the forecast data is not used for the estimation of the climatology, the function *Ano_CrossValid* automatically computes a climatology for each start date without taking that start date into account and subtracts it from the original data:

```
ano <- Ano_CrossValid(data$mod, data$obs)
```

3.3. Verification

To measure the skill of the two forecast systems, Pearsons correlation coefficient and the RMSE of the ensemble mean are computed and the results are plotted using *PlotVsLTime* (Figs. 5 and 6). As in *PlotClim*, the initial month can be specified with *monini*.

[4] This object can be downloaded directly from: www.bsc.es/projects/earthscience/s2dverification/s2dv_example_data.RData.

```
corr <- Corr(Mean1Dim(ano$ano_exp, 2),  # Mean1Dim allows to compute
             Mean1Dim(ano$ano_obs, 2))  # the ensemble mean.
PlotVsLTime(corr,
            monini = 11, freq = 12, ytitle = 'Correlation',
            listexp = c('Experiment A', 'Experiment B'),
            fileout = 'corr_expA_expB_obsX.jpeg')

rms <- RMS(Mean1Dim(ano$ano_exp, 2), Mean1Dim(ano$ano_obs, 2))
PlotVsLTime(rms,
            ytitle = 'Surface temperature (K)',
            monini = 11, freq = 12,
            listexp = c('Experiment A', 'Experiment B'),
            fileout = 'rms_expA_expB_obsX.jpeg')
```

In order to compute modes of variability and/or spatial correlations, we use two-dimensional fields. Using the *Load* function, the data sets are interpolated onto a common *T*106 Gaussian grid (the coarsest grid among the considered data sets) with CDO's distance-weighted method:

```
map_data <- Load('tas', list(expA, expB), list(obsX),
                 sdates = sdates,
                 leadtimemin = 1, leadtimemax = 12,
                 latmin = -10, latmax = 60,
                 lonmin = 100, lonmax = 250,
                 output = 'lonlat', grid = 't106grid',
                 method = 'conservative')
```

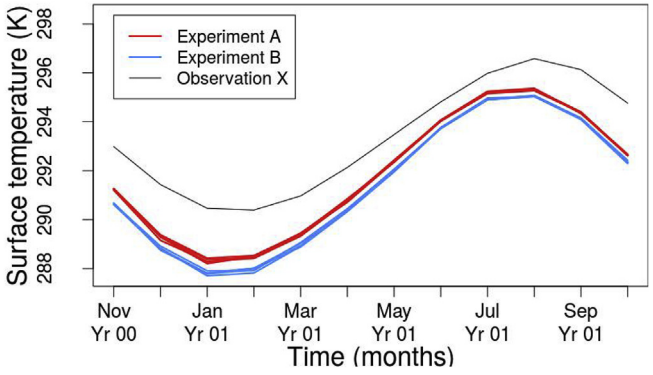

Fig. 3. Comparison of the two experiment climatologies. Each member's climatology is plotted in red (Experiment A) or blue (Experiment B). The reference climatology is drawn in black. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

The returned forecast and observational reference data now has two additional spatial dimensions[5]:

```
c(n. of data sets, n. of members, n. of start dates,
n. of forecast time steps, n. of latitudes, n. of longitudes)
map_ano <- Ano_CrossValid(map_data$mod, map_data$obs)
```

The modes of variability can be computed with the *EOF* function, which automatically applies area weighting to the input data. By default, the first 10 modes are provided. They can be plotted with *PlotEquiMap*. For example, the second mode of the reference shows a dipole between Alaska and northern Japan (Fig. 7).

[5] This R object can be downloaded directly from www.bsc.es/projects/earthscience/s2dverification/s2dv_example_map_data.RData.

```
eof_obs <- EOF(map_ano$ano_obs[1, 1, , 1, , ],  # The mode is computed for
               map_data$lon, map_data$lat)      # the first forecast time
                                                # step of the first member
                                                # and data set.
PlotLayout(PlotEquiMap, plot_dims = c('lat', 'lon'),
           eof_obs$EOFs[1:6, , ], map_data$lon, map_data$lat,
           brks = 41, bar_limits = c(-0.05, 0.05),
           titles = paste(signif(eof_obs$var, 1), '%'),
           bar_label_scale = 2,
           fileout = 'equimap_EOFs_obsX.jpeg')
```

The index of this mode can be computed for each ensemble member by projecting the anomaly map on the mode of variability, with *ProjectField*, and the results can be displayed with *PlotBoxWhisker*, as in Fig. 8. Reference indices are available in the *EOF* outputs.

```
mode1_exp <- ProjectField(map_ano$ano_exp, eof_obs, 1)
PlotBoxWhisker(mode1_exp[1, , , 1], eof_obs$PCs[, 1],
               ytitle = "Index", monini = 11, yearini = 1991,
               expname = 'Exp. A', obsname = 'Obs. X',
               fileout = '1st_pc_expA_obsX.jpeg')
```

Spatial correlations and confidence intervals at all forecast times and start dates, for all experiments, are computed via *ACC*. The results are plotted for Experiment A with *PlotACC*, as shown in Fig. 9.

```
acc <- ACC(map_ano$ano_exp, map_ano$ano_obs)
PlotACC(Subset(acc$ACC, 1, 1), sdates, freq = 12,
        ytitle = "Anomaly correlation coefficient",
        legends = c('Experiment A'),
        fileout = 'acc_expA_obsX.jpeg')
```

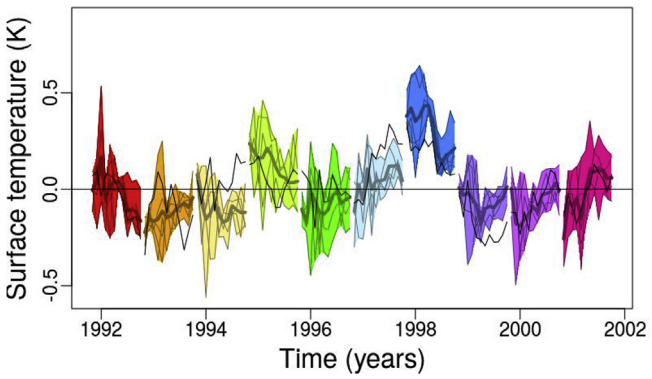

Fig. 4. Drift-corrected anomalies for Experiment A. The different colours represent different start dates. Each member appears with a light line and the ensemble mean with a bold line. The interval between maxima and minima is coloured. References appear in black. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

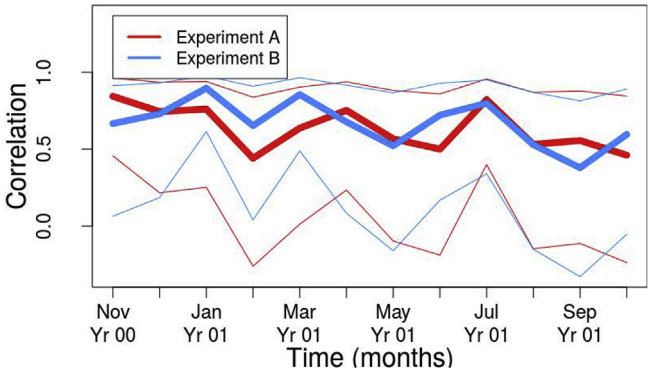

Fig. 5. Time correlation of the two experiments with the observational reference. The thick lines represent the correlation of the ensemble-mean with the references. The confidence intervals are plotted with a thinner line of the same colour.

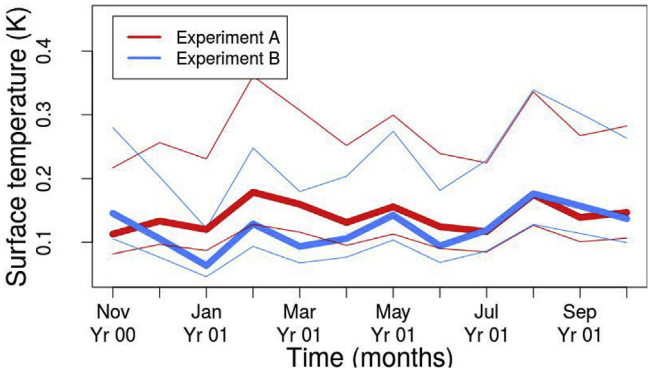

Fig. 6. RMSE for the two experiments with respect to the observational reference. The thick lines represent the RMSE of the ensemble-mean. The confidence intervals are plotted with a thinner line of the same colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

As well as the spatial correlation, maps of the temporal correlation can be plotted with the *PlotEquiMap* function as follows:

```
map_corr <- Corr(Mean1Dim(map_ano$ano_exp, 2),
                 Mean1Dim(map_ano$ano_obs, 2))
PlotEquiMap(map_corr[1, 1, 2, 1, , ], map_data$lon, map_data$lat,
            brks = 41, bar_limits = c(-1, 1),
            dots = map_corr[1, 1, 2, 1, , ] > map_corr[1, 1, 4, 1, , ],
            bar_label_scale = 2, intylat = 5, intxlon = 5,
            fileout = 'equimap_corr_raw_expA_obsX.jpeg')
PlotEquiMap(map_corr[2, 1, 2, 1, , ], map_data$lon, map_data$lat,
            brks = 41, bar_limits = c(-1, 1),
            dots = map_corr[2, 1, 2, 1, , ] > map_corr[2, 1, 4, 1, , ],
            bar_label_scale = 2, intylat = 5, intxlon = 5,
            fileout = 'equimap_corr_raw_expB_obsX.jpeg')
```

The results are shown in Fig 10 (*PlotStereoMap* could also have been used to produce the same plots but with a stereographic projection).

The ensemble forecast can be converted into probabilistic forecasts by binning and counting the forecast values. This conversion is done automatically in the *UltimateBrier* function, which can compute the standard Brier score (and its decomposition) after binning the forecasts into terciles:

We can see from Figs. 10 and 11 that both forecasts show similar spatial variability, however in general, Experiment A outperforms Experiment B, with the former having higher correlations and lower Briers scores for most regions. The region with the most skill in both experiments is $5^{\circ}S - 5^{\circ}N$, $190 - 240^{\circ}E$, also known as the Niño 3.4 region.

```
map_bs <- UltimateBrier(map_ano$ano_exp, map_ano$ano_obs, type = 'BS',
                        thr = c(1/3, 2/3), decomposition = FALSE)
PlotEquiMap(map_bs[1, 1, 1, 1, , ], map_data$lon, map_data$lat,
            brks = 41, bar_limits = c(0, 1), bar_label_scale = 2,
            color_fun = clim.palette('redblue'),
            intylat = 5, intxlon = 5,
            fileout = 'equimap_bs_expA_obsX.jpeg')
PlotEquiMap(map_bs[2, 1, 1, 1, , ], map_data$lon, map_data$lat,
            brks = 41, bar_limits = c(0, 1), bar_label_scale = 2,
            color_fun = clim.palette('redblue'),
            intylat = 5, intxlon = 5,
            fileout = 'equimap_bs_expB_obsX.jpeg')
```

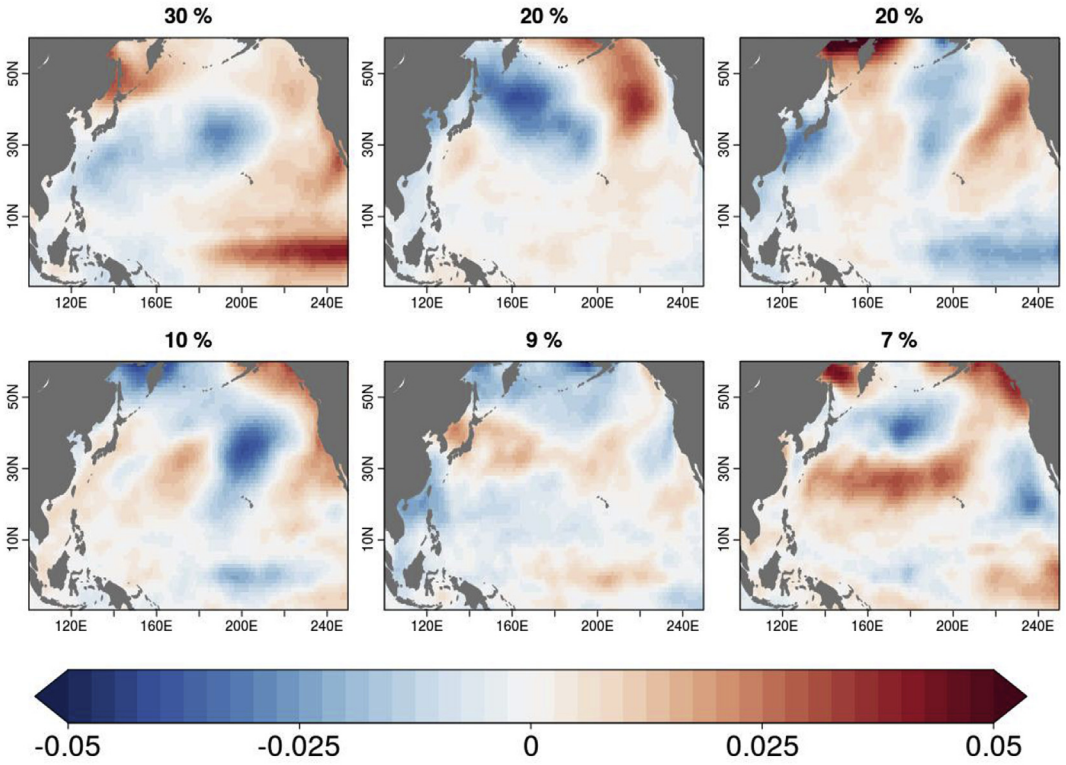

Fig. 7. Reference EOFs of the November North Pacific near surface temperature computed between 1991 and 2000. The numbers above the figures are the percentage of the variance explained by each mode.

4. Conclusions

This paper has introduced *s2dverification*, an R package for streamlining and simplifying the forecast verification process in climate research and other disciplines. Its key features have been described, including a powerful data loading and homogenization function, functions that transparently implement well known methods for pre-processing and verifying climate data, novel verification and synthetic forecast generation methods, and functions to generate frequently required visual products.

The stages of data retrieval, pre-processing, computation of scores and visualisation of data and results have been briefly described and put into context, together with an explanation of the typical challenges encountered and the corresponding mechanisms and functions the package provides to deal with these challenges, and a short review of the relevant technical aspects.

Given the proliferation of verification packages, it seems likely

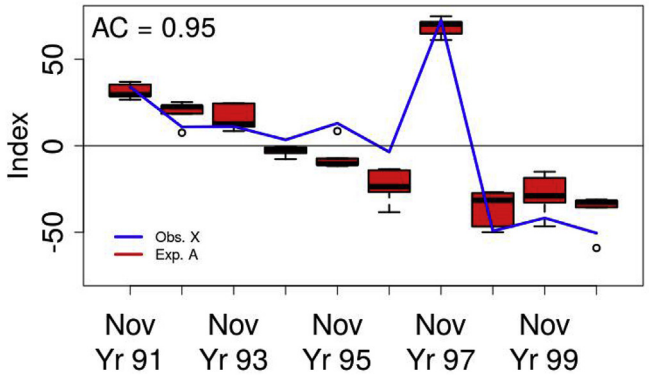

Fig. 8. Box-and-whisker plot of the first mode of variability of Experiment A. The red boxes show the distribution of the indices of the first mode for each start date in November with the median in black, the quartiles are given by the upper and lower limits of the boxes and outliers are shown as points. The blue curve shows the reference index. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

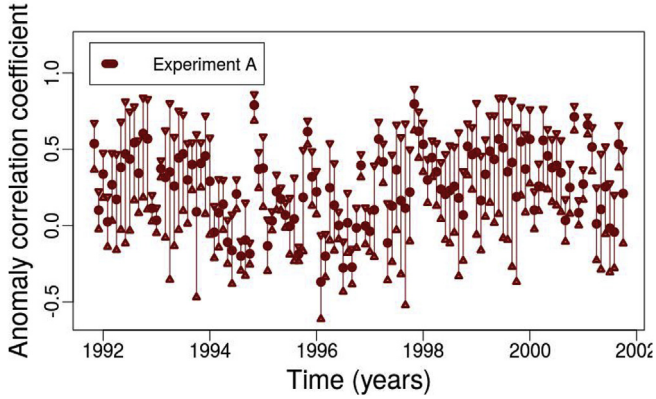

Fig. 9. Spatial anomaly correlation coefficients for Experiment A against the reference. The forecast values from each experiment are plotted at intervals of one month and the vertical lines show the confidence intervals.

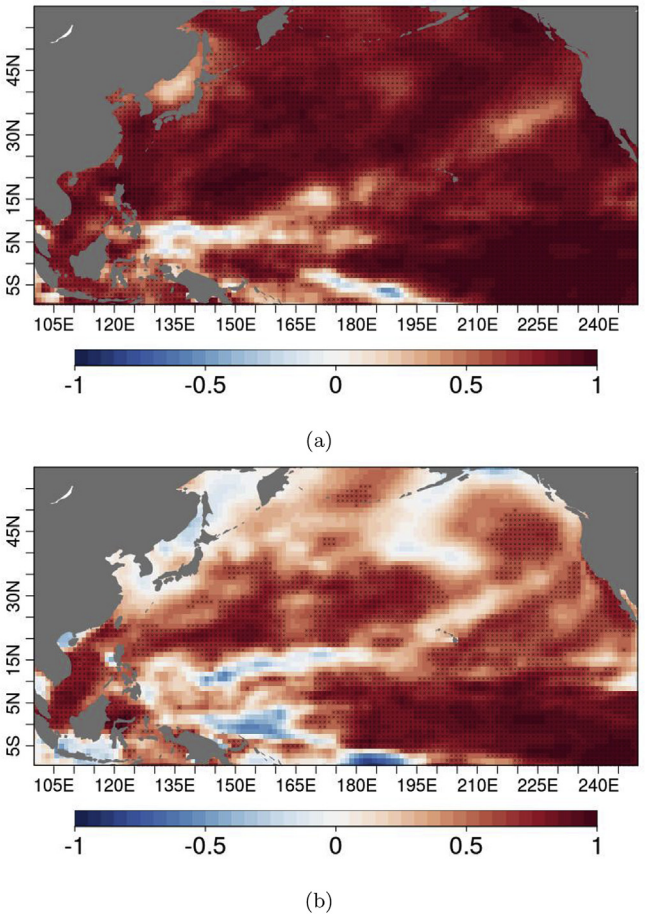

Fig. 10. Correlation between a) Experiment A and b) Experiment B with the reference anomalies, computed at each grid point over the North Pacific. A black dot indicates the correlation is significant at the 5% significance level.

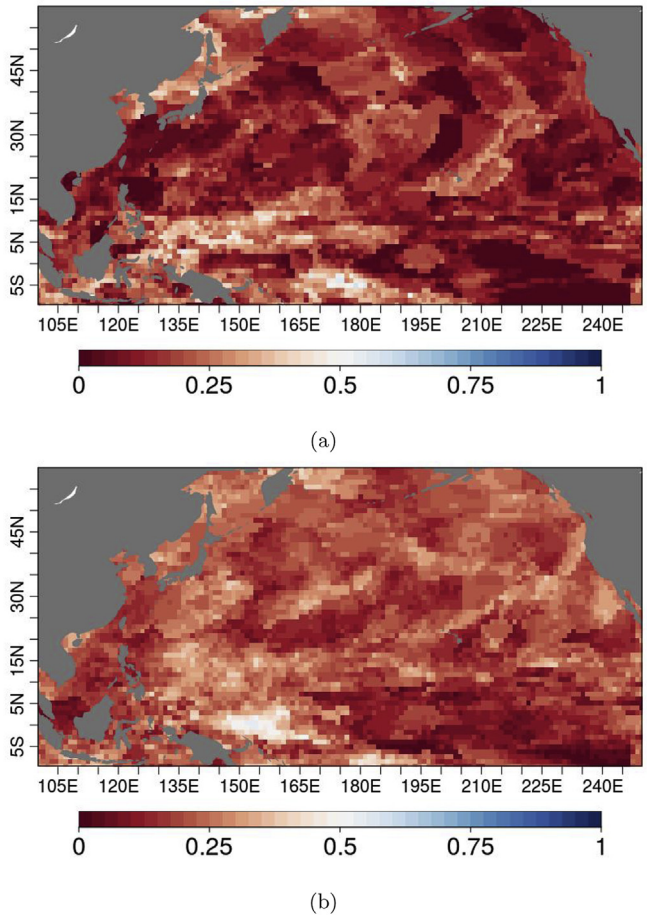

Fig. 11. Plot of the Brier score for November of a) Experiment A and b) Experiment B against Observation X, at each grid point over the North Pacific.

(and desirable) that their developers will have to make a concerted effort to converge in the near future, either by agreeing on a set of conventions and good practices to make the various tools compatible with all forecast data structures or by creating a single scalable and sustainable verification framework.

There are several further developments planned for *s2dverification*. The data retrieval module is currently being extended to support data sets stored in additional file schemes and additional file formats, and with temporal resolutions other than daily or monthly, as well as data sets with start dates at the sub-daily level. Also, the verification scores and pre-processing functions are being adapted to work on multi-core platforms.

For the mid-term future, the package is being extended to trace the steps followed from the beginning to the end of the verification process to ensure the reproducibility of results. In terms of visualisation, the time series and map plotting functions are being unified and extended with additional options. Finally, the overall usability of the framework is being improved to make some technical details more transparent to the user, and to efficiently handle large forecast data sets distributed on cluster platforms, taking into consideration existing solutions in *raster*, *sp*, *zoo*, *ff*, and other packages.

Acknowledgements

The research leading to these results has received funding from the EU Seventh Framework Programme (FP7, 2007–2014) under grant agreements 308378 (SPECS), 607085 (EUCLEIA), 603521 (PREFACE) and 308291 (EUPORIAS); and from the Copernicus Climate Change Services (C3S) contract 2016/C3S_51_LOT3_BSC/SC1(QA4Seas). O. Bellprat's contract has been

financed by the European Space Agency under The Living Planet Fellowship, as part of the project VERITAS-CCI.

Software availability

Software s2dverification
Description The *s2dverification* software is freely available as an R package. A comprehensive framework for forecast verification. Functions are provided for loading and processing the forecast and reference data, calculating verification statistics and visualizing the results
Main developers V. Guemas and N. Manubens
Source language R
Availability https://cran.r-project.org/web/packages/s2dverification/index.html

Appendix A. Verification statistics

Throughout this section, x_i represents the forecast value for reference y_i, for $i = 1, \ldots, N$ total references.

Root-mean-square error (RMSE): the square root of the mean of the squared differences. The RMSE is a measure of the distance between the forecast and the references, i.e. the forecast accuracy.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \tag{1}$$

Root-mean-square skill score (RMSSS): a skill score based on the RMSE. The RMSSS compares the RMSE of the forecasts with that of the climatology at a fixed location.

$$RMSSS = \frac{100}{N}\sum_{i=1}^{N}\left[1 - \sqrt{\frac{(\widehat{x}_i - x_i)^2}{(\overline{x} - x_i)^2}}\right] \tag{2}$$

Brier score: the mean square differences between the forecast probability x_i and reference probability y_i. The BS takes values between 0 and 1 with lower values corresponding to more accurate forecasts. For further information, and a decomposition of the Brier score see Ferro and Fricker (2012).

$$BS = \sum_{i=1}^{N}(x_i - y_i)^2. \tag{3}$$

Appendix B. Use case directory tree

In the case study the data files are distributed as depicted in the following directory tree:

```
|--experiments/
|    |--experimentA/
|    |    |--monthly_mean/
|    |         |--tas/
|    |              |--tas_19911101.nc
|    |              |--tas_19921101.nc
|    |              |        .
|    |              |        .
|    |              |        .
|    |              |--tas_20001101.nc
|    |--experimentB/
|         |--monthly_mean/
|              |--tas/
|                   |--tas_19911101.nc
|                   |--tas_19921101.nc
|                   |        .
|                   |        .
|                   |        .
|                   |--tas_20001101.nc
|--observations/
     |--observationX/
          |--monthly_mean/
               |--tas/
                    |--tas_199101.nc
                    |--tas_199102.nc
                    |--tas_199103.nc
                    |--tas_199104.nc
                    |        .
                    |        .
                    |        .
                    |--tas_200108.nc
                    |--tas_200109.nc
                    |--tas_200110.nc
```

These files need to be NetCDF 3/4 compliant and contain the variable 'tas', fulfilling the guidelines detailed in *?Load*.

References

Appelhans, T., Detsch, F., Nauss, T., et al., 2015. Remote: empirical orthogonal teleconnections in R. J. Stat. Software 65, 1–19.

Balmaseda, M.A., Mogensen, K., Weaver, A.T., 2013. Evaluation of the ECMWF ocean reanalysis system ORAS4. Q. J. R. Meteorol. Soc. 139, 1132–1161.

Bellprat, O., Doblas-Reyes, F., 2016. Attribution of extreme weather and climate events overestimated by unreliable climate simulations. Geophys. Res. Lett. 43, 2158–2164.

Benestad, R.E., Mezghani, A., Parding, K.M., 2016. Esd: Climate Analysis and Empirical-statistical Downscaling (ESD) Package for Monthly and Daily Data. http://rcg.gvc.gu.se/edu/esd.pdf.r.package.version.1.2.

Bennett, N.D., Croke, B.F., Guariso, G., Guillaume, J.H., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T., Norton, J.P., Perrin, C., et al., 2013. Characterising performance of environmental models. Environ. Model. Software 40, 1–20.

Bhend, J., 2016. EasyVerification: Ensemble Forecast Verification for Large Data Sets. http://www.meteoswiss.ch.rpackage.version0.2.0.

Bretherton, C.S., Smith, C., Wallace, J.M., 1992. An intercomparison of methods for finding coupled patterns in climate data. J. Clim. 5, 541–560.

Brodeau, L., Barnier, B., Treguier, A.-M., Penduff, T., Gulev, S., 2010. An ERA40-based atmospheric forcing for global ocean circulation models. Ocean Model. 31, 88–104.

Brown, B.G., Gotway, J.H., Bullock, R., Gilleland, E., Fowler, T., Ahijevych, D., Jensen, T., 2009. The model evaluation tools (MET): community tools for forecast evaluation. In: Preprints, 25th Conf. On International Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Phoenix, AZ, Amer. Meteor. Soc. A (P. 6). Volume 9.

Brown, J.D., Demargne, J., Seo, D.-J., Liu, Y., 2010. The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environ. Model. Software 25, 854–872.

Brownrigg, R., 2013. Package Maps. https://cran.r-project.org/web/packages/maps/index.html.

Caron, L.-P., Boudreault, M., Camargo, S.J., 2015. On the variability and predictability of eastern Pacific tropical cyclone activity. J. Clim. 28, 9678–9696.

Caron, L.-P., Jones, C.G., Doblas-Reyes, F., 2014. Multi-year prediction skill of Atlantic hurricane activity in CMIP5 decadal hindcasts. Clim. Dynam. 42, 2675–2690.

Deckmyn, A., 2015. Package Mapproj. https://cran.r-project.org/web/packages/mapproj/index.html.

Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., et al., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137, 553–597.

Déqué, M., 2003. Deterministic forecasts of continuous variables. Forecast Verification: A Practitioner's Guide in Atmospheric Science, Second Edition, pp. 77–94.

Dijkstra, H.A., 2013. Nonlinear Climate Dynamics. Cambridge University Press.

Doblas-Reyes, F.J., García-Serrano, J., Lienert, F., Biescas, A.P., Rodrigues, L.R., 2013. Seasonal climate predictability and forecasting: status and prospects. Wiley Interdisciplinary Reviews: Climatic Change 4, 245–268.

Du, H., Doblas-Reyes, F., García-Serrano, J., Guemas, V., Soufflet, Y., Wouters, B., 2012. Sensitivity of decadal predictions to the initial atmospheric and oceanic perturbations. Clim. Dynam. 39, 2013–2023.

Dussin, R., Barnier, B., Brodeau, L., 2014. The Making of Drakkar Forcing Set DFS5. DRAKKAR/MyOcean Rep. 05-10, vol. 14.

Emanuel, K., 2005. Increasing destructiveness of tropical cyclones over the past 30 years. Nature 436, 686–688.

Eyring, V., Righi, M., Evaldsson, M., Lauer, A., Wenzel, S., Jones, C., Anav, A., Andrews, O., Cionni, I., Davin, E., et al., 2015. ESMValTool (v1. 0)–a community diagnostic and performance metrics tool for routine evaluation of Earth System Models in CMIP. Geosci. Model Dev. Discuss. (GMDD) 8, 7541–7661.

Ferro, C., 2014. Fair scores for ensemble forecasts. Q. J. R. Meteorol. Soc. 140, 1917–1923.

Ferro, C.A., Fricker, T.E., 2012. A bias-corrected decomposition of the Brier score. Q. J. R. Meteorol. Soc. 138, 1954–1960.

Fučkar, N.S., Guemas, V., Johnson, N.C., Massonnet, F., Doblas-Reyes, F.J., 2016. Clusters of interannual sea ice variability in the northern hemisphere. Clim. Dynam. 47, 1527–1543.

Fučkar, N.S., Volpi, D., Guemas, V., Doblas-Reyes, F.J., 2014. A posteriori adjustment of near-term climate predictions: accounting for the drift dependence on the initial conditions. Geophys. Res. Lett. 41, 5200–5207.

García-Serrano, J., Doblas-Reyes, F., 2012. On the assessment of near-surface global temperature and North Atlantic multi-decadal variability in the ENSEMBLES decadal hindcast. Clim. Dynam. 39, 2025–2040.

García-Serrano, J., Losada, T., Rodríguez-Fonseca, B., Polo, I., 2008. Tropical Atlantic variability modes (1979-2002). part ii: time-evolving atmospheric circulation related to sst-forced tropical convection. J. Clim. 21, 6476–6497.

Gilleland, E., Gilleland, M.E., 2016. Package spatialvx. Framework 15, 51–64.

Group, S.M., 2015. downscaleR: Climate Data Manipulation and Statistical Downscaling. https://github.com/SantanderMetGroup/downscaleR/wiki r package version 0.8-2.

Guemas, V., Auger, L., Doblas-Reyes, F., Rust, H., Ribes, A., 2014. Dependencies in statistical hypothesis tests for climate time series. Bull. Am. Meteorol. Soc. 95, 1666–1668.

Hijmans, R.J., van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J.A., Lamigueiro, O.P., Bevan, A., Racine, E.B., Shortridge, A., et al., 2015. Package raster. R package.

Hurrell, J., 1996. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. Oceanogr. Lit. Rev. 2, 116.

Kharin, V., Boer, G., Merryfield, W., Scinocca, J., Lee, W.-S., 2012. Statistical adjustment of decadal predictions in a changing climate. Geophys. Res. Lett. 39.

Kousky, V.E., Kagano, M.T., Cavalcanti, I.F., 1984. A review of the Southern Oscillation: oceanic-atmospheric circulation changes and related rainfall anomalies. Tellus 36, 490–504.

Krishnamurti, T., Rajendran, K., Vijaya Kumar, T., Lord, S., Toth, Z., Zou, X., Cocke, S., Ahlquist, J.E., Navon, I.M., 2003. Improved skill for the anomaly correlation of geopotential heights at 500 hpa. Mon. Weather Rev. 131, 1082–1102.

Lees, J.M., 2012. Geomapdata: Data for Topographic and Geologic Mapping. http://CRAN.R-project.org/package=geomapdata r package version 1.0-4.

Meehl, G.A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., Corti, S., Danabasoglu, G., Doblas-Reyes, F., Hawkins, E., et al., 2014. Decadal climate prediction: an update from the trenches. Bull. Am. Meteorol. Soc. 95, 243–267.

Mendlik, T., Heinrich, G., Gobiet, A., Leuprecht, A., 2015. From climate model ensembles to statistics: introducing the" wux" package. In: EGU General Assembly Conference Abstracts, vol. 17, p. 12266.

Potts, J.M., 2003. Basic Concepts. Forecast Verification: a Practitioner's Guide in Atmospheric Science, Second Edition, pp. 11–29.

R Core Team, 2015. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria. http://www.R-project.org/.

Rasmusson, E.M., Wallace, J.M., 1983. Meteorological aspects of the El Nino/southern oscillation. Science 222, 1195–1202.

Schulzweida, U., 2015. CDO 2015: Climate Data Operators. http://www.mpimet.mpg.de/cdo.

Siegert, S., 2015. SpecsVerification: Forecast Verification Routines for the SPECS FP7 Project. http://CRAN.R-project.org/package=SpecsVerification r package version 0.4-1.

Siegert, S., Stephenson, D.B., Sansom, P.G., Scaife, A.A., Eade, R., Arribas, A., 2015. A Bayesian Framework for Verification and Recalibration of Ensemble Forecasts: How Uncertain is NAO Predictability? arXiv preprint arXiv:1504.01933.

Slingo, J., Palmer, T., 2011. Uncertainty in weather and climate prediction. Phil. Trans. Roy. Soc. Lond. Math. Phys. Eng. Sci. 369, 4751–4767.

Smith, D.M., Cusack, S., Colman, A.W., Folland, C.K., Harris, G.R., Murphy, J.M., 2007. Improved surface temperature prediction for the coming decade from a global climate model. Science 317, 796–799.

Stephenson, D.B., Coelho, C.A., Jolliffe, I.T., 2008. Two extra components in the Brier score decomposition. Weather Forecast. 23, 752–757.

Stephenson, D.B., Pavan, V., Bojariu, R., 2000. Is the North Atlantic oscillation a random walk? Int. J. Climatol. 20, 1–18.

Trenberth, K.E., 1984. Some effects of finite sample size and persistence on meteorological statistics. part ii: potential predictability. Mon. Weather Rev. 112, 2369–2379.

Vecchi, G.A., Zhao, M., Wang, H., Villarini, G., Rosati, A., Kumar, A., Held, I.M., Gudgel, R., 2011. Statistical-dynamical predictions of seasonal North Atlantic hurricane activity. Mon. Weather Rev. 139, 1070–1082.

Venegas, S.A., 2001. Statistical methods for signal detection in climate. Danish Center for Earth System Science Report 2, 96.

Villarini, G., Vecchi, G.A., 2012. North Atlantic power dissipation index (PDI) and accumulated cyclone energy (ACE): statistical modeling and sensitivity to sea surface temperature changes. J. Clim. 25, 625–637.

Villarini, G., Vecchi, G.A., Smith, J.A., 2010. Modeling the dependence of tropical storm counts in the North Atlantic basin on climate indices. Mon. Weather Rev. 138, 2681–2705.

Villarini, G., Vecchi, G.A., Smith, J.A., 2012. US landfalling and North Atlantic hurricanes: statistical modeling of their frequencies and ratios. Mon. Weather Rev. 140, 44–65.

Von Storch, H., Zwiers, F.W., 2001. Statistical Analysis in Climate Research. Cambridge university press.

Wallace, J.M., Lim, G.-H., Blackmon, M.L., 1988. Relationship between cyclone tracks, anticyclone tracks and baroclinic waveguides. J. Atmos. Sci. 45, 439–462.

Weigel, A., Liniger, M., Appenzeller, C., 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q. J. R. Meteorol. Soc. 134, 241–260.