Verification score card

Maria-Helena Ramos, Jeffrey Norville, Guillaume Thirel, Florian Pappenberger, Ilias Pechlivanidis

imprex

Learn from today to anticipate tomorrow

IMPROVING PREDICTIONS AND MANAGEMENT OF HYDROLOGICAL EXTREMES

~

Verification score card

(Deliverable 4.1)

@imprex_eu

Funded by
the Horizon 2020
Framework Programme
of the European Union

Grant agreement No. 641811

Deliverable	Deliverable title
Related Work Package:	WP4 – Improved predictability of hydrological extremes
Deliverable lead:	IRSTEA
Author(s):	Maria-Helena Ramos (IRSTEA), Jeffrey Norville (IRSTEA), Guillaume Thirel (IRSTEA), Florian Pappenberger (ECMWF), Ilias Pechlivanidis (SMHI)
Contact for queries	maria-helena.ramos@irstea.fr
Grant Agreement Number:	n° 641811
Instrument:	HORIZON 2020
Start date of the project:	01.10.2015
Duration of the project:	48 months
Website:	www.IMPREX.eu
Abstract	This report presents the prototype of a verification score card (hereafter called, scoreboard) for hydrological reforecasts developed for IMPREX. Based on a review of verification scores for hydrological forecasts and scoreboard displays, we propose a utility that combines a score database and a web-based display to facilitate the visualization and interpretation of the quality of the forecasts produced during the project. An initial testing on reforecast data is also presented to illustrate the way scores are displayed by the utility developed.

Dissemination level of this document

X	PU	Public
	PP	Restricted to other programme participants (including the Commission Services)
	RE	Restricted to a group specified by the consortium (including the European Commission Services)
	CO	Confidential, only for members of the consortium (including the European Commission Services)

Versioning and Contribution History

Version	**Date**	**Modified by**	**Modification reasons**
v.01	21/07/2016	MH Ramos	Introduction of Section 1
v.02	25/07/2016	MH Ramos	Introduction of Section 2 and 3.1
v.03	26/07/2016	J. Norville	Introduction of Section 3.2 to 3.5
v.04	27/07/2016	G. Thirel, J. Norville	Revision of Section 3
v.05	29/07/2016	MH Ramos	Introduction of Section 4. Final revision from co-authors. Version sent to quality-check review
v.06	02/08/2016	MH Ramos	Minor revision of typos after quality-check review
v.1	11/10/2016	MH Ramos	Revised (final) version including remarks from Bart van den Hurk (KNMI, coord.).
v.2	26/06/2017	MH Ramos	Final version after remarks from EU Project Officer and reviewers, and considering update of the scoreboard display

Table of Contents

IMPREX has received funding from the European Union Horizon 2020 Research and Innovation Programme under Grant agreement N° 641811

List of Figures

List of Tables

Glossary

Forecast verification: the process of comparing forecasts to relevant observations, or assessing the quality of a forecast product or a forecast system.

Forecast quality: how well a forecast compares against a corresponding observation of what actually occurred, or some good estimate of the true outcome.

Forecast value: how a forecast helps the user to make a better decision.

Forecast skill: the relative accuracy of the forecast over some reference forecast.

Score: a quantitative measure of forecast quality.

Skill score: a relative measure (or scaled representation) of forecast quality that relates the forecast accuracy of a particular forecast to some reference forecast. Skill scores range from negative infinity to positive one. A perfect categorical forecast yields skill values of 1. A forecast with similar skill to the reference forecast will have a skill score of zero, and a forecast that is less skilful than the reference forecast will have negative skill score values.

Scoreboard utility: a graphical interface, connected to a score database, to support comparisons between numerical scores of different forecasts or forecast systems.

Score data provider: a partner of the IMPREX project who contributes data to the scoreboard.

Scoreboard user: any person (partner of the IMPREX project or not) who wishes to visualize the quality of a forecast or a forecast system investigated in one of the case-studies of the IMPREX project (under the condition that the score of this forecast or a forecast system is made available by a score data provider).

1 Introduction

One of the main aims of the IMPREX project is to enhance the forecast quality of extreme hydro-meteorological conditions and, as a result, their impacts. Impacts are measured for six targeted types of users involved in the following socio-economic sectors: Floods, Hydropower, Agriculture and droughts, Navigation, Water supply, and Water economy. This wide range of users also represents a wide range of needs concerning the performance of a forecast system. For instance, while a user in the flood protection sector may be more interested in the moment the level of a river or stream will exceed a critical flood alert level, a user in the hydropower sector may search to improve the quality of the weekly or monthly averages of river inflows to a water reservoir to plan its operation.

It was considered useful to the IMPREX project to have a common framework to evaluate and inter-compare the performance of different forecast systems over the case study applications of the project. This report presents **a prototype of a verification scoreboard for hydrological reforecasts,** developed to be used during the IMPREX project. Our focus is placed on the main technical aspects of the implementation of the scoreboard utility and its capabilities. The aim is to introduce the technical concepts behind the tool in order to better satisfy the needs of its potential score data providers (namely, the project's partners implementing the case studies).

This report is organized as follows: In **Section 2**, we present an overview of the main aspects involved in the verification of hydrological forecasts. The aim is to highlight the main issues one has to have in mind when developing a scoreboard. **Section 3** focuses on the utility that was developed, which combines a score database and a web-based display to facilitate the visualization and the interpretation of the quality of the forecasts produced during the project. An initial testing on reforecast data is presented to exemplify the approach adopted. **Section 4** provides a brief conclusion and proposes ways forward for the use and maintenance of the utility during the lifetime of the project and beyond.

2 Verification of hydrological forecasts

2.1 A brief overview of forecast verification

A large part of the performance of a forecast system is measured by the quality of the forecasts delivered by the system. The quality of a forecast measures how well a forecast is compared against a corresponding observation of what actually occurred, or some good estimate of the true outcome. Forecast verification is the process of assessing the quality of a forecast. In hydrology, forecast verification is also called "forecast evaluation". In meteorology, the "value of a prediction" specifically refers to how a forecast helps the user make better decisions.

The quality of a forecast can be assessed through a wide range of metrics (or verification scores)[1] (see details in Wilks, 2011; Jolliffe and Stephenson, 2012). Many forecasters and practitioners recommend the use of several scores to better assess the attributes of a forecast (such as reliability, resolution, discrimination and sharpness): "*any set of forecasts can then be ranked as best, second best,..., worst, according to a chosen score, though the ranking need not be the same for different choices of score.*"[2]

What scores to choose depends on the user's objectives as well as on the characteristics of the forecasts being evaluated. In several situations, it may be interesting to evaluate the relative quality of a forecast with regard to a reference system or a baseline (e.g., to decide which is better between system A and system B). Here, again, there exist several baselines that can be considered, varying from simple approaches such as climatology (always forecasting an average value) or persistence (the last observation is forecast to persist into the future) to more sophisticated benchmarks that take into account analogue-based features (see, for instance, the twenty-three benchmarks that were designed and used for the assessment of hydrological forecasts in the study proposed by Pappenberger *et al.*, 2015).

[1] See, for instance, http://www.cawcr.gov.au/projects/verification/

[2] from: Jolliffe, I. T., and D. B. Stephenson (2012), 2nd Edition, Chapter 1 Introduction, page 5.

Relative measures of quality are known as *skill scores*. They measure the improvement of a given forecast relative to a reference forecast, or how much better one forecasting system is in comparison to another forecasting system given the same variables and location.

There are several reasons for assessing the quality of a forecast dataset. Measures of forecast quality can be used, for instance, to guide decisions on the additional human and/or financial resources needed to further develop a forecasting system; to judge how changes in modelling approaches affect the quality of the forecasts; to improve our understanding of uncertainties and biases involved in the forecast processes; and to understand the drivers of forecast performance in a modelling system as well as the strengths and weakness of different forecasting approaches.

Forecast quality may depend on the forecast situation or conditions (e.g., setup of the forecast verification framework). It also varies with human and modelling factors, forecast lead time, forecast location, as well as temporal and spatial scales considered[3]. Providing information on these factors and on the calculations of the verification scores is therefore important in the communication of forecast quality assessments.

2.2 Main aspects of the verification of hydrological forecasts

Verification of hydrological model outputs (river flows or water levels) is a continuing and necessary effort for modellers and operational forecasters to provide guidance on model development, implementation and on the operational use of hydrological forecasts. Many of the metrics used in the verification of hydrological forecasts come from the meteorological community or have been adapted to answer to the specific needs encountered in hydrological analyses (Casati *et al.*, 2008; Pappenberger *et al.*, 2008).

[3] http://hepex.irstea.fr/hepex-science-and-challenges-verification-of-ensemble-forecasts/

Most often, typical constraints of the modelling of hydrological processes require metrics that target a given variable or focus on a specific aspect of quality searched by an advanced forecast user (Demargne *et al.*, 2009; Brown *et al.*, 2010): for instance, metrics have been developed to evaluate how accurate predictions of flood peaks are in magnitude and timing (e.g., Liu *et al.*, 2011; Zappa *et al.*, 2013) or to evaluate highly temporally correlated seasonal low flows (e.g., Wood *et al.*, 2005; Nicolle *et al.*, 2014; Trambauer *et al.*, 2015).

Hydrological forecast verification focuses more and more on quantifying the quality of probabilistic or ensemble-based predictions (Cloke and Pappenberger, 2009), although metrics for deterministic forecasts are also commonly used (e.g., correlation coefficient, Nash–Sutcliffe model efficiency coefficient or Persistence Index), as well as traditional metrics for categorical events associated to contingency tables (e.g. probability of detection, false alarm rate or Critical success index). A variety of metrics exist for a variety of forecast goals, variables to be evaluated, associated space and time steps, etc. Examples of this diversity of applications of verification metrics to evaluate forecasts in hydrological studies are presented in Table A1.1 in **Annex 1**.

2.3 Particularities within the IMPREX project

The IMPREX project focuses on enhancing forecast quality of extreme hydro-meteorological conditions and their impacts. Its partners work to develop methods and tools to improve the forecasting of meteorological and hydrological extremes. It is then important to quantitatively measure if (and how) improved forecasts perform better than the reference forecasts or the benchmark system that is being improved. Additionally, it is also of interest to measure how meteorological forecast improvements impact hydrological forecasts, as well as how hydrological forecast improvements impact risk assessment and management variables, following the flow of information and models that is typically seen in an end-to-end hydro-meteorological forecasting chain. In IMPREX, the value of the improved forecast information is demonstrated in a set of case studies, which evaluate hydrological risks and impacts that are relevant to stakeholders at the regional and European scales. These case studies cover six water sectors, as shown below.

1. Central European Rivers: floods, transport, agriculture and droughts

2. Bisagno river basin in Italy: floods
3. Thames River Basin in UK: floods
4. South-East French Catchments: hydropower
5. Júcar River Basin in eastern Spain: hydropower, agriculture and droughts
6. Lake Como basin in the Italian Alpine region: hydropower, agriculture and droughts
7. Upper River Umeälven in Sweden: hydropower
8. Segura River Basin in the Iberian Peninsula: urban water, agriculture and droughts
9. The Llobregat River Basin in north-eastern Spain: urban water
10. The Messara valley in Crete: agriculture and droughts

Many evaluations of forecast quality will be carried out along the different case studies and sectors during the lifetime of the IMPREX project. Side by side evaluations would help partners and stakeholders to better understand the impacts of the forecasts on the sectors investigated. To facilitate inter-comparisons, a verification scoreboard for the hydrological reforecasts is proposed.

The scoreboard should allow partners to easily store their score data, computed for the many configurations of improved forecasts they may want to test, and visualize the evolution of scores with lead time, across locations within a case study, and using different metrics of forecast quality. In IMPREX, a verification scoreboard should include short, medium, and long ranges. It should also be flexible enough to accept metrics that are traditionally used in forecast verification as well as new metrics, specifically developed to focus on hydrological variables of special interest (e.g., hydrologic threshold exceedances, hydrological extremes of high and low flows) and to quantify the added value of improved forecasts for specific applications (e.g., gain in lead time from anticipating drought conditions, gain in energy production from the use of better forecasts in models of hydropower reservoir optimization, evaluation of transport costs by using probabilistic forecasts in a simulation model for river navigation, etc.).

3 Design of a verification scoreboard utility for IMPREX

In order to contribute to the forecast verification goals of the IMPREX project, a prototype of a verification scoreboard was designed. A standard data format, database and graphical interface composing this utility are presented in the next sections, after a brief overview of examples of existing scoreboards.

3.1 Examples of existing verification scoreboards

Communicating forecast verification data or scores is an important aspect of an operational forecasting centre. It allows forecast developers to be transparent on the quality of their forecasts and users to gain confidence on the forecasts they are using when making decisions. There are several ways to communicate verification statistics. Usually, forecast centres show on their websites, from their operational or the research activities, charts, graphs, numerical tables or summarized information on the quality of past forecasts or on the verification results of studies carried out for specific important events. In general, forecast verification displays are accompanied by instructions on how scores are computed and should be interpreted.

Some examples of scoreboards can be found from the websites of operational forecasting centres. They are shown in **Annex 2**. For instance, Figure A2.1 shows a screenshot of the webpage on verification statistics from the Weather Prediction Centre of the National Weather Service of NOAA in USA. The example shows the evolution of the Bias and the Threat scores over the period 1970-2015 for different lead times. From the same source, Figure A2.2 shows the verification of Quantitative Precipitation Forecasts (QPF) provided for a single event: the Hurricane Sandy, which affected most of the eastern United States (especially the coastal Mid-Atlantic States), in the autumn 2012. Verification is provided through the comparison of the maps of the accumulated precipitation forecasts and the accumulated observed precipitation over the affected areas.

Another example can be found in the website of the UK MetOffice for global long-range predictions. Probabilistic skill maps and plots are available and updated monthly for temperature and rainfall predictions up to six months ahead. Figure A2.3 illustrates how the user can change the “skill score type” to display a ROC score map or a Reliability diagram.

Other options on the visual display include the variable to display, the geographic area, and the period used for the computation of the scores.

The World Meteorological Organization (WMO) has also proposed a verification board for its long-range forecast verification system, jointly managed by the Australian Bureau of Meteorology and the Meteorological Service of Canada. Figure A2.4 shows the display choices available from this board, which include the model to be verified, the type of score, lead time, period, etc. The chart catalogue for forecast verification proposed by ECMWF is illustrated in Figure A2.5. The large variety of options for display and visualizing a given score as a function of several parameters (lead time, variable, threshold, period of computation, etc.) is highlighted.

It is out of scope of this deliverable to achieve the same level of display and generality of the scoreboards examples presented here. Our goal is to propose a utility that can be used during the lifetime of the project to centralize forecast verification scores on hydrological reforecasts for the water sectors studied in the IMPREX project. Scoreboards for hydrological forecasts are rare and this prototype is an attempt to face the challenge of adapting graphic and display options for forecast verification focusing on hydrological variables. The design of the score database and the setup of the display interface are presented in details in the next sections.

3.2 Overview of the concept behind the IMPREX scoreboard utility

Sharing a scoreboard – a tool which allows comparison and collaboration – is an excellent opportunity for a web page, specifically if it is a combination of a back-end data repository with a graphical, shared, interface. This allows analyses and updates to be performed by several users, potentially at the same time. For the IMPREX project, we developed a database (which can reside on a centralized server) as the repository, and an HTML5 graphical user interface (GUI) front-end to query and display score data. **Figure 1** illustrates the verification scoreboard concept set up for the IMPREX project.

A user can connect to the scoreboard utility (App) and start a query for a given case (pertaining to a case study of IMPREX), selecting the choices related to the score the user wants to visualize. This will make the interface display R plots and tables, which the user can export as a PDF file. The interface is connected to the IMPREX score database, which will be fed with scores computed by the IMPREX partners involved in the evaluation of the forecasts for the IMPREX case studies.

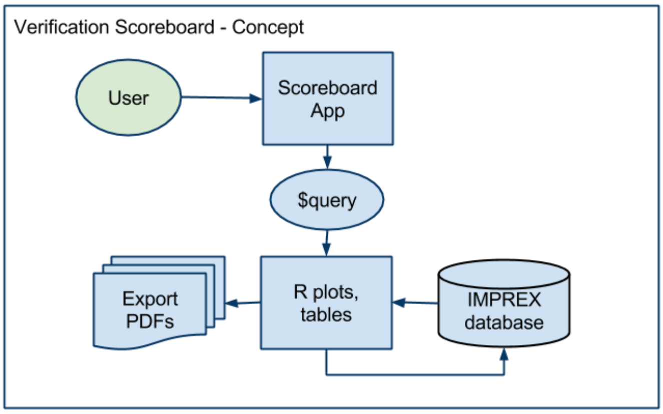


Figure 1: Workflow diagram of the verification scoreboard

3.3 Design of the score database

The scoreboard was developed on open-source tools R and PostgreSQL. R is a programming environment and language popularly used for statistical programming. PostgreSQL is a robust database system based on SQL, structured query language. We have incorporated an R package developed by RStudio called "Shiny" to produce the web-based interactive GUI.

In the future, it can be envisaged that the scoreboard may be accessed two ways:

1. By navigating to the IMPREX website and clicking on the link to the server URL;
2. By downloading or cloning the source repository to run the application through a local R installation (additional libraries may be needed).

While the first option is more anticipated, and the best way to connect to the shared IMPREX score database, a local installation allows users to experiment with file upload formats, visualizing their score input files locally, and creating local databases as needed.

The user may also use the local scoreboard without an internet connection, for example during a presentation.

3.3.1 SQL database: PostgreSQL

During the development phase, we used a combination of local database instances[4] at partner IRSTEA (France) and an instance running on Amazon's free AWS service and local database instances.

Local installers for the PostgreSQL database are available for Linux, Mac and Windows platforms (see **Table 1** for software specifications). A GUI interface (e.g. "pgAdmin") is typically installed along with the instance of the database to help manipulate the database, set password information, and backup data. To clone the schema of the IMPREX database, a DDL (Data Definition Language) script is available, along with more detailed instructions for first-time users.

Adding or connecting to multiple databases is facilitated by an environment text- or ini-file, ".Renviron". User information (username, password) is pulled from this file for local database instances. A template for this file is included in **Annex 3**.

Table 1: Software specification for postgreSQL

Installation source:	https://www.postgresql.org/download/
Version:	Tested on 9.4 and 9.5

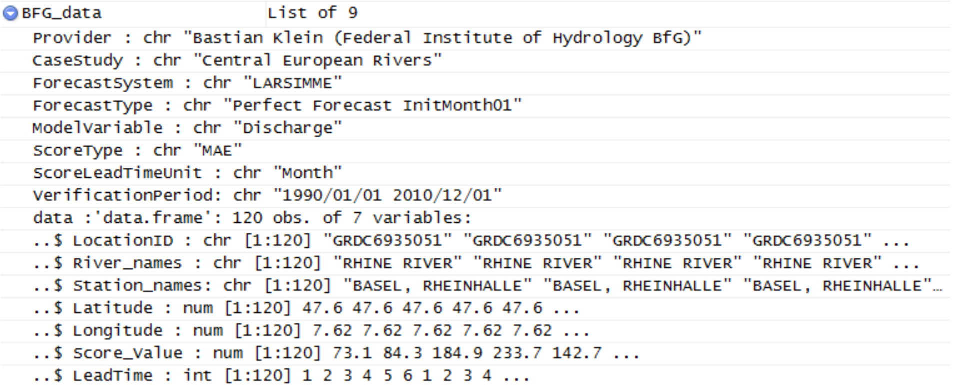
[4] A database is a collection of files that reside on a server. A database instance is a set of memory structures that manage database files. The instance can also refer to the allocated memory and collection of processes running on a server.

3.3.2 Format of data files

In the schema proposed in Figure 1, the "IMPREX database" has score data that need to be provided by the project partners in a pre-defined way. In order to match project requirements, we have identified the RDS R data file format[5] as the optimal data file format supported.

Compared to the more commonly-used RData file format, RDS prevents from possibly erasing internal variables of the application if a variable name is identical to an object name in an imported file. From a user perspective, producing an RDS file is very similar to producing an RData file. The scoreboard utility also supports text file imports, although it is important to note that file size limitations may play a role on uncompressed files.

Data files support the following common data elements on any user-imported score file: 1. Metadata; 2. Data. The metadata contains contact information for the individual most responsible for the data submittal (an IMPREX partner), as well as details such as model verification start and end dates, score type, case study and forecast system details (see **Figure 2** for an illustration). Scores are included last, and take the biggest part of the file space. **Annex 4** presents a file specification and further explanations.



```
BFG_data                  List of 9
  Provider : chr "Bastian Klein (Federal Institute of Hydrology BfG)"
  CaseStudy : chr "Central European Rivers"
  ForecastSystem : chr "LARSIMME"
  ForecastType : chr "Perfect Forecast InitMonth01"
  Modelvariable : chr "Discharge"
  ScoreType : chr "MAE"
  ScoreLeadTimeUnit : chr "Month"
  VerificationPeriod: chr "1990/01/01 2010/12/01"
  data :'data.frame': 120 obs. of 7 variables:
  ..$ LocationID : chr [1:120] "GRDC6935051" "GRDC6935051" "GRDC6935051" "GRDC6935051" ...
  ..$ River_names : chr [1:120] "RHINE RIVER" "RHINE RIVER" "RHINE RIVER" "RHINE RIVER" ...
  ..$ Station_names: chr [1:120] "BASEL, RHEINHALLE" "BASEL, RHEINHALLE" "BASEL, RHEINHALLE"…
  ..$ Latitude : num [1:120] 47.6 47.6 47.6 47.6 47.6 ...
  ..$ Longitude : num [1:120] 7.62 7.62 7.62 7.62 7.62 ...
  ..$ Score_Value : num [1:120] 73.1 84.3 184.9 233.7 142.7 ...
  ..$ LeadTime : int [1:120] 1 2 3 4 5 6 1 2 3 4 ...
```

Figure 2: Example of the header of a score data file (metadata)

[5] R has its own data file format and usually uses the .rds extension. The .rds file format is usually smaller than its text file counterpart and will therefore take up less storage space. The .rds file will also preserve data types such as factors and dates, eliminating the need to redefine data types after loading the file.

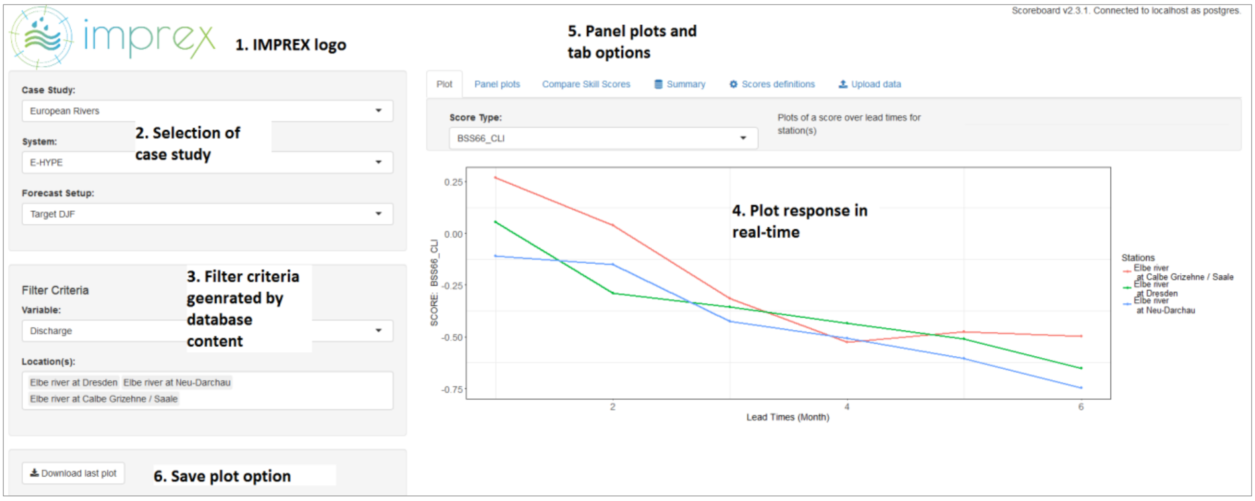
3.4 Design of the scoreboard interface

The layout of the scoreboard features typical web page navigation tools; though R software is used to create the webpages, they are standard HTML5 with JavaScript and readable by any platform and browser.

3.4.1 HTML display

The web page was developed using R, and specifically RStudio's Shiny package (see **Table 2** for software specifications). **Figure 3** illustrates the main components of the HTML display. We have added the IMPREX logo on the top left corner. In the space below, the left column is dedicated to allow the user to select the IMPREX case study and the system for which scores are to be visualized. A number of filter choices are available, so that the user can choose the location to display (e.g. sub-basins of a catchment included in a case study or location of point stations included in the European-wide case studies), the variable (e.g. streamflow or precipitation), the forecast system (e.g., the benchmark system, a system where bias correction was applied), and the score to plot.

The graphical plot is displayed in real-time, once all the choices have been made. The tabs in the upper part of the plot area allow the user to visualize panel plots (i.e., several score plots for inter-comparison purposes), compare skill scores, and access a summary table of the database, the scores definition and an upload tool. Plots and table can be saved in a PDF file and downloaded.



Scoreboard v2.3.1. Connected to localhost as postgres.

Case Study: European Rivers

System: E-HYPE

Forecast Setup: Target DJF

Filter Criteria

Variable: Discharge

Location(s): Elbe river at Dresden Elbe river at Neu-Darchau Elbe river at Calbe Grizehne / Saale

Score Type: BSS66_CLI

Plots of a score over lead times for station(s)

1. IMPREX logo

2. Selection of case study

3. Filter criteria geenrated by database content

4. Plot response in real-time

5. Panel plots and tab options

6. Save plot option

Figure 3: Example of the HTML display of the scoreboard utility

Table 2: Software specification for R and RStudio

Browser compatibility	Tested on Ubuntu Chrome, Firefox; Windows Chrome, Explorer, Edge, Firefox
Installation source, R:	https://cran.r-project.org/
Version:	Tested on 3.3.1 and 3.4.0
Installation source, RStudio:	https://www.rstudio.com/products/rstudio/download/
Version:	Tested on 0.99.879 and 1.0.143

3.4.2 Workflow

As shown in Figure 3, once on the scoreboard webpage, the user finds a screen containing selection boxes on the left column, plots on the right column and navigation tabs above the plot. The two approaches to reduce the dataset shown (and work with dynamic plots) are to: 1. Select a listed Case Study, 2. Select a System.

This selection updates the remaining navigation filters, and presents the user with a narrower set of choices. The list of items available in the filter criteria is thus adapted according to the available items in the Score database, which are automatically detected by the interface once the selection of Case Study and System has been done. The user can:

1. select one (or more) location(s), a list of locations appear and update the plots,
2. select one score to view the associated plot (in the main plot area).

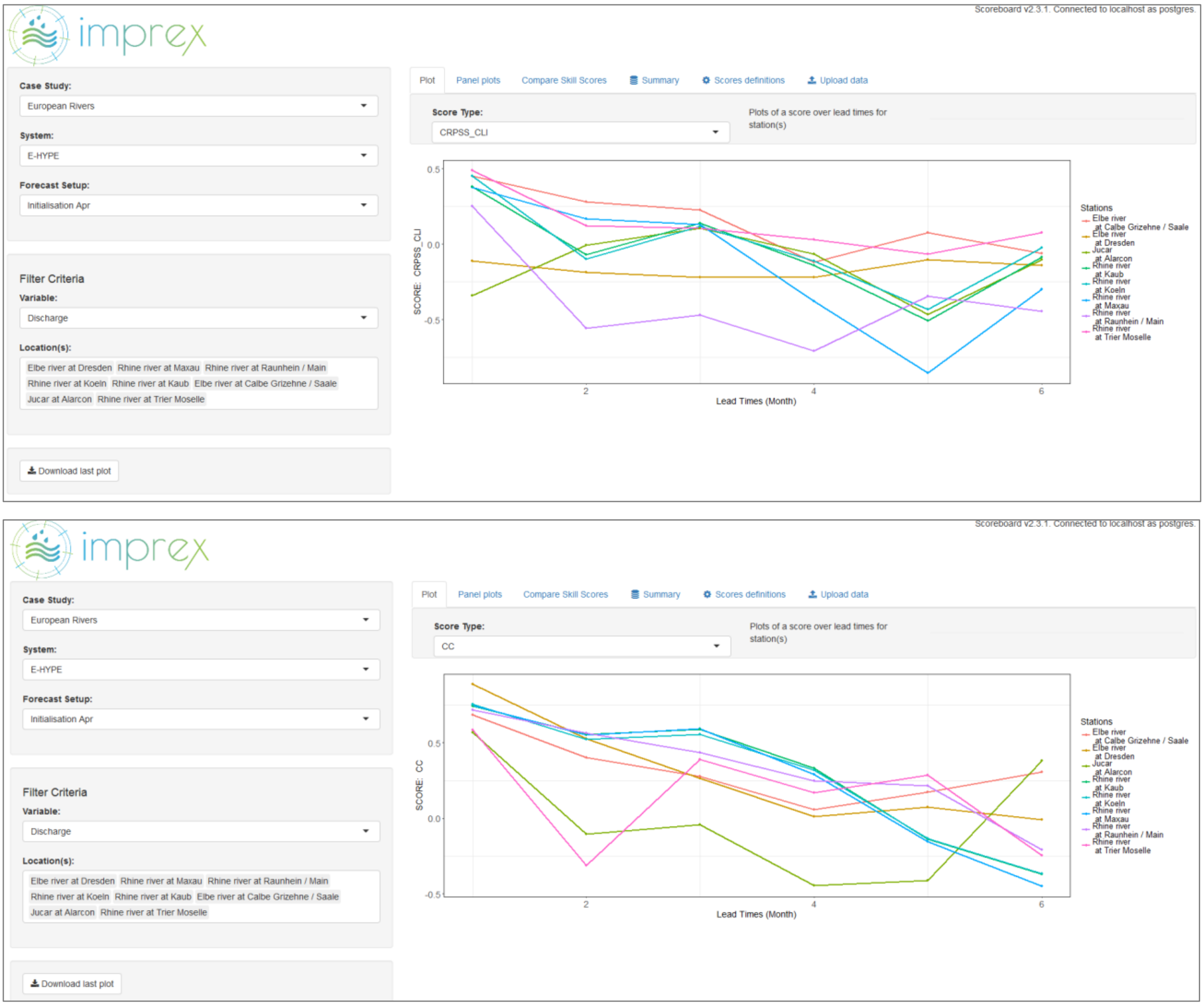
At any point during the use of the Scoreboard the user may view a panel plot with all scores available for the selected locations, download a PDF of the last plot drawn, as well as view a summary table of the database.

An alternative workflow may suit more experienced R users, or those testing their data import format. If the user has installed RStudio (and necessary libraries) locally, they may clone the Scoreboard project to their disk and run it with the following command: shiny::runApp. While this Scoreboard cannot connect to the IMPREX database nor make submissions to it, it can accept and verify a user's RDS or txt file uploads, as well as working with and visualizing the data locally.

3.5 Example with a test data from IMPREX partner

The Scoreboard utility was tested based on a test dataset provided by partner SMHI. Data comes from streamflow simulations based on the E-HYPE seasonal forecasting system in Europe, which were verified by SMHI with a variety of scores. These E-HYPE score data were provided in RData file format. It contained scores and skill scores for several stations in Europe. The file contained scores for each month of the year, 6 lead months and each station.

Figure 4 and **Figure 5** show examples of the visualization obtained from the scoreboard utility for score plots, panel plots and summary table, and for selected locations and scores in the E-HYPE test score dataset.



Case Study: European Rivers

System: E-HYPE

Forecast Setup: Initialisation Apr

Filter Criteria

Variable: Discharge

Location(s): Elbe river at Dresden Rhine river at Maxau Rhine river at Raunhein / Main Rhine river at Koeln Rhine river at Kaub Elbe river at Calbe Grizehne / Saale Jucar at Alarcon Rhine river at Trier Moselle

Score Type: CRPSS_CLI

Case Study: European Rivers

System: E-HYPE

Forecast Setup: Initialisation Apr

Filter Criteria

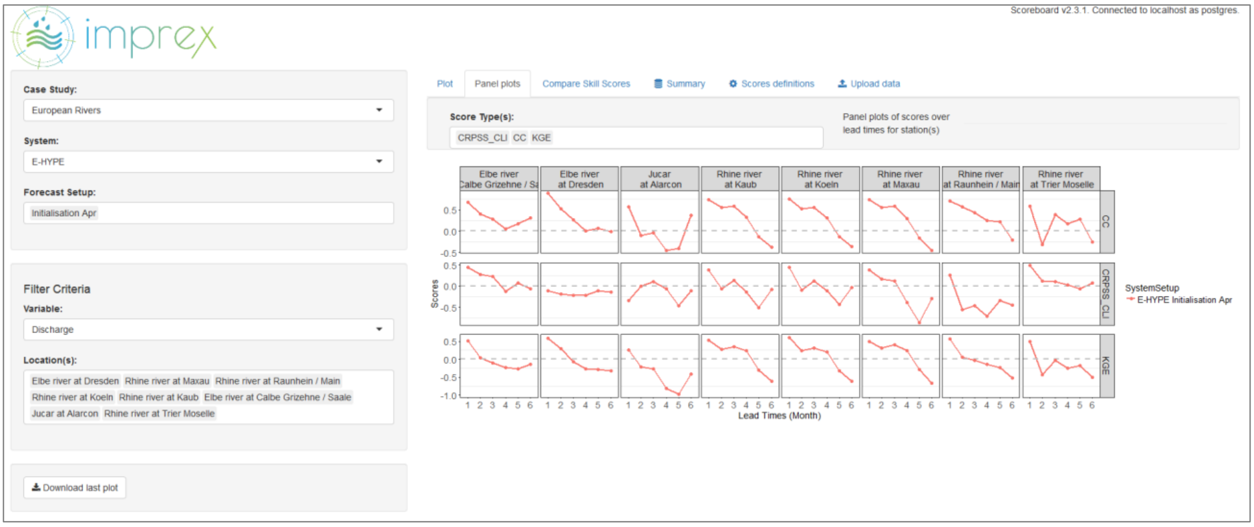
Variable: Discharge

Location(s): Elbe river at Dresden Rhine river at Maxau Rhine river at Raunhein / Main Rhine river at Koeln Rhine river at Kaub Elbe river at Calbe Grizehne / Saale Jucar at Alarcon Rhine river at Trier Moselle

Score Type: CC

Figure 4: Screenshots of the HTML display interface of the scoreboard utility for 8 locations in the E-HYPE test score dataset, the CRPSS score (top) and Correlation coefficient (bottom)

IMPREX has received funding from the European Union Horizon 2020 Research and Innovation Programme under Grant agreement N° 641811



Scoreboard v2.3.1. Connected to localhost as postgres.

Case Study: European Rivers

System: E-HYPE

Forecast Setup: Initialisation Apr

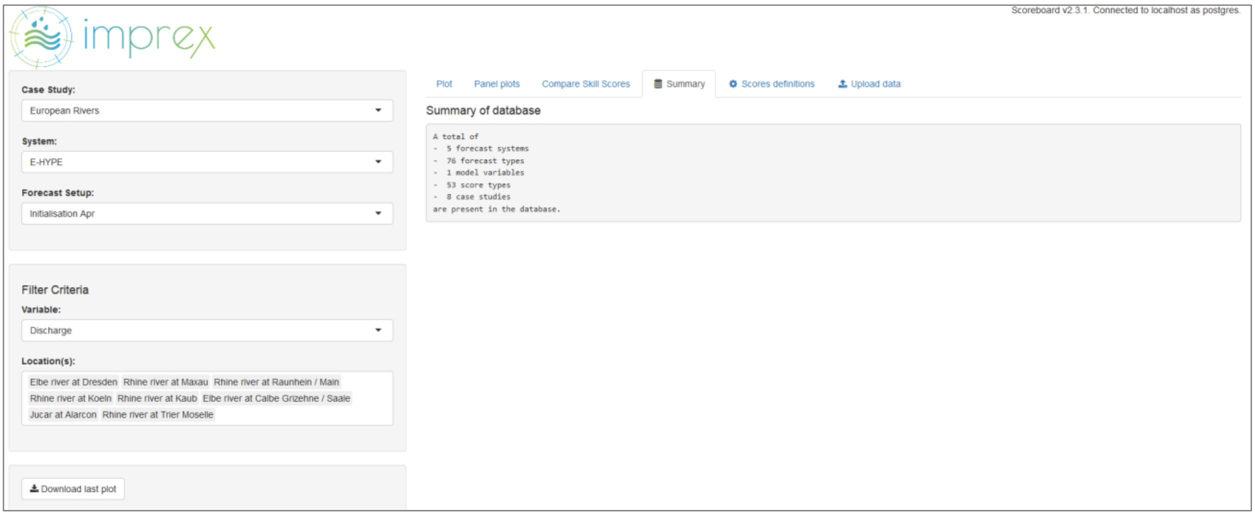
Filter Criteria

Variable: Discharge

Location(s): Elbe river at Dresden, Rhine river at Maxau, Rhine river at Raunhein / Main, Rhine river at Koeln, Rhine river at Kaub, Elbe river at Calbe Grizehne / Saale, Jucar at Alarcon, Rhine river at Trier Moselle

Score Type(s): CRPSS_CLI CC KGE

Panel plots of scores over lead times for station(s)



Scoreboard v2.3.1. Connected to localhost as postgres.

Case Study: European Rivers

System: E-HYPE

Forecast Setup: Initialisation Apr

Filter Criteria

Variable: Discharge

Location(s): Elbe river at Dresden, Rhine river at Maxau, Rhine river at Raunhein / Main, Rhine river at Koeln, Rhine river at Kaub, Elbe river at Calbe Grizehne / Saale, Jucar at Alarcon, Rhine river at Trier Moselle

Summary of database

```
A total of
-  5 forecast systems
-  76 forecast types
-  1 model variables
-  53 score types
-  8 case studies
are present in the database.
```

Figure 5: Screenshots of the tab options of the HTML display interface of the scoreboard utility: panel plots (top) with 3 scores (CRPS Skill score, Correlation coefficient and KGE) and 8 locations, and summary table of the E-HYPE test score dataset (bottom)

4 Conclusions and way forward

We have developed and tested a prototype of verification scoreboard utility for the IMPREX project. It is based on a database as the repository for score data collected during the lifetime of the project and an HTML5 graphical user interface (GUI) front-end to query the database and display plots of the scores that were uploaded to it. Following the development and first test phase, priorities and needs were discussed with IMPREX partners, allowing to enhance the utility.

We proposed a RDS format file for the score data, with possibility to a txt file format, after data import process had been agreed among partners. We have written the specifications for the data format and a loader utility was included to facilitate IMPREX partners to upload their score data files.

Some issues remaining include:

- A variety of types of plots can be found in the literature. We have included some of the more common plots. IMPREX partners may have other suggestions and these could eventually be implemented, if technically feasible, including some more integrated and easy-to-understand plots for a general public of users of the scoreboard.
- We have left the possibility in the format file to collect location data (geographic coordinates). The display can thus be enhanced with maps of locations and/or scores.
- We have worked with typical statistical scores, but other types of scores (including those measuring economic gains in the water sectors investigated in IMPREX) could also be included.
- An important step is to identify the host for the database and the web application (Shiny server). This issue was discussed during the 2nd Imprex General Assembly and a decision was taken together with IMPREX partners, the website administrator and coordinator to explore a URL link between partners ECMWF and ARCTIK.
- User manuals with detailed descriptions of the scoreboard utility can be provided to introduce the score data providers and the users to the tool.

5 References

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson D., Salamon P., 2014: Evaluation of ensemble streamflow predictions in Europe. *J. Hydrol.*, 517, 913–922.

Alvarez-Garreton, C., D. Ryu, A. W. Western, C.-H. Su, W. T. Crow, D. E. Robertson, C. Leahy, 2015: Improving operational flood ensemble prediction by the assimilation of satellite soil moisture: comparison between lumped and semi-distributed schemes. *Hydrol. Earth Syst. Sci.*, 19, 1659–1676.

Bourgin, F., M.-H., Ramos, G., Thirel, V., Andréassian, 2014: Investigating the interactions between data assimilation and post-processing in hydrological ensemble forecasting, *J. Hydrol.*, Vol. 519, 2775–2784.

Brown, J. D., Demargne, J., Seo, D.J., and Liu, Y., 2010: The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environ. Modell. Softw.*, 25(7), 854-872.

Brown, J.D., He, M., Regonda, S., Wu, L., Lee, H., Seo, D.-J, 2014: Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification. *J. Hydrol.*, Vol. 519, 2847-2868.

Carpenter, T. M., and K. P. Georgakakos, 2004: Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model. *J. Hydrol.*, 298, 202–221.

Casati, B., Wilson, L.J., Stephenson, D.B, 2008: Forecast verification: current status and future directions, *Meteorological Applications*, 15 (1), 3–18.

Cloke, H., and Pappenberger, F., 2009: Ensemble flood forecasting: a review. *J. Hydrol.*, 375 (3–4), 613–626.

Demargne J., M. Mullusky, K. Werner, T. Adams, S. Lindsey, N. Schwein, W. Marosi, and E. Welles, 2009: Application of Forecast Verification Science to Operational River Forecasting in the U.S. National Weather Service. *Bull. Amer. Meteor. Soc.*, 90 (6): 779-784.

Fan, F.M., Collischonn, W., Meller, A., Botelho, L.C.M., 2014: Ensemble streamflow forecasting experiments in a tropical basin: The São Francisco river case study. *J. Hydrol.*, Vol. 519, 2906-2919.

Franz, K.J., Hogue, T.S., Barik, M., 2014: Assessment of SWE data assimilation for ensemble streamflow predictions. *J. Hydrol.*, Vol. 519, Part D, 2737–2746.

Hashino, T., Bradley, A.A., Schwartz, S.S., 2007: Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrol. Earth Syst. Sci.*, 11, 939-950.

Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* Second Edition. Wiley and Sons, New York. doi: 10.1002/9781119960003

Liechti, K., Zappa, M., Fundel, F., Germann, U., 2013: The potential of radar-based ensemble forecasts for flash-flood early warning in the southern Swiss Alps. *Hydrol. Earth Syst. Sci.*, 17, 3853–3869.

Liu, Y., J. D. Brown, J. Demargne, and D-J. Seo, 2011: A wavelet-based approach to assessing timing errors in hydrologic predictions. *J. Hydrol.*, 397 (3-4), 210-224.

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Le Lay, M., Besson, F., Soubeyroux, J.-M., Viel, C., and co-authors, 2014: Benchmarking hydrological models for low-flow simulation and forecasting on French catchments, *Hydrol. Earth Syst. Sci.*, 18, 2829-2857, doi:10.5194/hess-18-2829-2014

Pappenberger, F., K. Scipal, and R. Buizza, 2008: Hydrological aspects of meteorological verification. *Atmos. Sci. Lett.*, 9, 43–52.

Pappenberger, F., Ramos, M.H., Cloke, H.L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., Salamon, P., 2015: How do I know if my forecasts are better? Using benchmarks in Hydrological Ensemble Prediction, *J. Hydrol.*, 522: 697-713.

Roulin, E., Vannitsem, S., 2015: Post-processing of medium-range probabilistic hydrological forecasting: Impact of forcing, initial conditions and model errors. *Hydrol. Process.*, Vol. 29, Issue 6, 1434-1449.

Trambauer, P., M. Werner, H. C. Winsemius, S. Maskey, E. Dutra, and S. Uhlenbrook, 2015: Hydrological drought forecasting and skill assessment for the Limpopo River basin, southern Africa. *Hydrol. Earth Syst. Sci.*, 19, 1695–1711, doi:10.5194/hess-19-1695-2015

Voisin, N., F. Pappenberger, D. P. Lettenmaier, R. Buizza, and J.C. Schaake, 2011: Application of a Medium-Range Global Hydrologic Probabilistic Forecast Scheme to the Ohio River Basin. *Wea. Forecasting*, 26, 425–446.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences : An Introduction*, Academic Press, 676 pages.

Wood, A. W., A. Kumar, and D. P. Lettenmaier, 2005: A retrospective assessment of National Centers for Environmental Prediction climate model–based ensemble hydrologic forecasting in the western United States. *J. Geophys. Res. Atmospheres*, 110, D04105, doi:10.1029/2004JD004508

Zappa, M., F. Fundel, and S. Jaun, 2013: A ‘Peak-Box’ approach for supporting interpretation and verification of operational ensemble peak-flow forecasts. *Hydrol. Process.*, 27 (1), 117–131, DOI: 10.1002/hyp.9521

6 Annexes

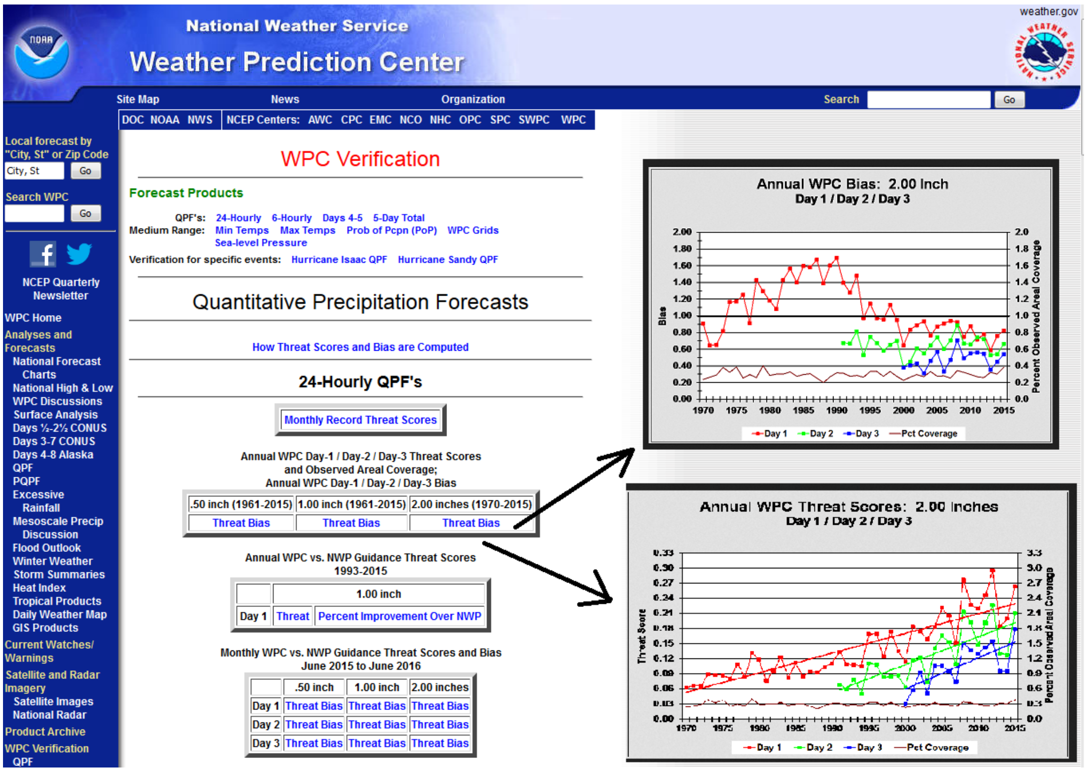
6.1 Annex 1 – Examples of applications of forecast verification metrics in hydrology

Geographic context, spatial domain, variable	Metrics used	Verification period	Verification against
To evaluate the assimilation of satellite soil moisture retrievals into a rainfall-runoff model for flood prediction in a large, sparsely monitored catchment (Alvarez-Garreton et al., 2015):			
1 semi-arid river basin in Queensland, Australia. Daily streamflow, lead time 0 (simulation)	Nash–Sutcliffe efficiency, RMSE, Peak volume error, Rank histogram, POD, FAR, CRPS	1 June 2003 – 2 March 2014	Streamflow records
To investigate the potential of radar-based ensemble flash-flood forecasts, including evaluation against deterministic discharge forecasts (Liechti et al., 2013):			
3 catchments in the southern Swiss Alps. Hourly runoff, lead time up to eight hours	Brier skill score, FAR and POD, ROC area	1389 hourly time steps (June 2007–Dec. 2010)	Observed discharge
To diagnose how rainfall input and parametric uncertainty influence flow simulation uncertainty in a distributed hydrologic model (Carpenter and Georgakakos, 2004):			
5 watersheds and sub-catchments in the southern Central Plains of the United States. Hourly flow simulations, lead time 0 (simulation)	Ensemble dispersion through a normalized inter-quantile range (90th and 10th percentile)	25 to 30 flow events for each watershed between June 1993 and May 1999	Observed streamflow
To evaluate the benefits of using ensemble predictions for reservoir inflow to hydropower plants, in comparison to the deterministic values given by the control member of the ensemble and by the ensemble mean. (Fan et al., 2014):			
São Francisco river basin and sub-catchments, in Minas Gerais, Southeast Brazil. Hourly streamflow, lead time up to 16 days	MAE, CRPS, Rank histogram, ROC Curve, Brier Skill Score, visual inspection, threshold exceedance diagrams for flood events	Three wet seasons in 2010-2013 and three selected major flood events	Observed hydrographs and inflows estimated by water balance.
To assess the impact of data assimilation for ESP seasonal water supply (Franz et al., 2014):			
North Fork of the American River Basin (NFARB) in northern California, USA. Water supply values (total discharge, m^3)	RMSE, Percent bias, Correlation coefficient, CRPSS, Containing ratio, Discrimination diagram, Reliability diagram	26 to 58 years of historic data for January to April,	Discharges and SWE observations
To evaluate bias correction methods for ensemble streamflow volume forecasts (Hashino et al., 2007):			
Des Moines River basin, in Iowa, north-central USA. Monthly flow volumes, issued sequentially for each month, lead times of 1 to 12 months	Mean square error (MSE) skill score using climatology as a reference,	1949 to 1996 historic data (forecasts generated on the first of each month)	Observed streamflow

To establish a baseline for future enhancements, and to guide the operational use of the ensemble forecasting system studied (Brown et al., 2014):			
8 catchments in USA. Daily averages of streamflows, lead times 1 to 14 days, and time aggregated discharges.	Relative mean error (RME) of the ensemble mean, correlation coefficient, CRPS, BSS, and decompositions, Reliability diagram, ROC	Hindcasts for a 20-year period between 1979 -1999.	Simulated and observed streamflows
To evaluate the EFAS (European Flood Awareness System) operational suite (Alfieri et al., 2014):			
38452 grid points of the EFAS European river network. Daily streamflows, lead times up to 10 days	Nash–Sutcliffe efficiency, Forecast bias, Coefficient of variation of the RMSE, CRPSS	Operational forecasts and hindcasts from 2009 -2012	A reference discharge simulation using observed meteorological data
To test a global approach to producing hydrological ensemble forecasts in river basins where in situ data are sparse (Voisin et al., 2011):			
4 outlets at the Ohio River basin, USA. Daily gridded runoff forecasts, lead times up to 15 days.	Bias, RMSE, Pearson correlation, Rank histograms, CRPSS	2002-2007	A reference discharge simulation using observed meteorological data
To investigate the impact of errors in the forcing, in the model structure and parameters, and in the initial conditions on hydrological forecasts; to test the post-processing of hydrological ensembles (Roulin and Vannitsem, 2015):			
The Ourthe Orientale at Mabompré in the Ardennes region in Belgium. Daily discharges, lead time up to 10 days	Bias or mean error (ME), RMSE, Spread, CRPS and decomposition	March 2008 to December 2012	Observed discharges and reference discharge simulation using observed meteorological data
To investigate how data assimilation and post-processing contribute to the skill of hydrological ensemble forecasts (Bourgin et al., 2014):			
202 catchments in France. Hourly streamflows, lead times up to 48 hours	Bias, RMSE, PIT, Normalized mean interquartile range (NMIQR), CRPSS	2005–2009	Observed discharges

6.2 Annex 2 – Examples of existing verification scoreboards

Screenshot of the Weather Prediction Centre verification webpage (National Weather Service, NOAA: http://www.wpc.ncep.noaa.gov/html/hpcverif.shtml) :



WPC Verification

Forecast Products

QPF's: 24-Hourly 6-Hourly Days 4-5 5-Day Total
Medium Range: Min Temps Max Temps Prob of Pcpn (PoP) WPC Grids Sea-level Pressure
Verification for specific events: Hurricane Isaac QPF Hurricane Sandy QPF

Quantitative Precipitation Forecasts

How Threat Scores and Bias are Computed

24-Hourly QPF's

Monthly Record Threat Scores

Annual WPC Day-1 / Day-2 / Day-3 Threat Scores
and Observed Areal Coverage;
Annual WPC Day-1 / Day-2 / Day-3 Bias

.50 inch (1961-2015)	1.00 inch (1961-2015)	2.00 inches (1970-2015)
Threat Bias	Threat Bias	Threat Bias

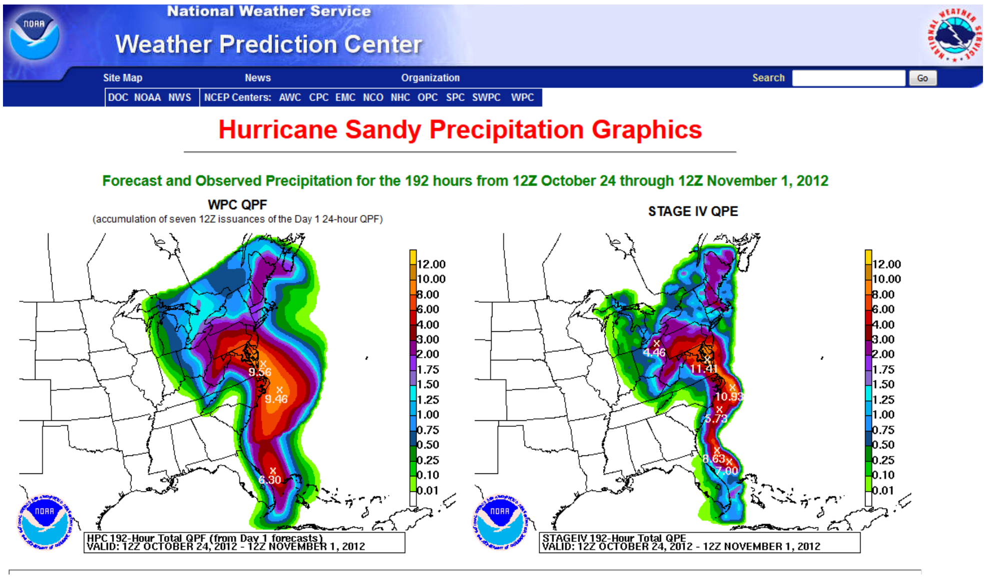
Annual WPC vs. NWP Guidance Threat Scores
1993-2015

	1.00 inch	
Day 1	Threat	Percent Improvement Over NWP

Monthly WPC vs. NWP Guidance Threat Scores and Bias
June 2015 to June 2016

	.50 inch	1.00 inch	2.00 inches
Day 1	Threat Bias	Threat Bias	Threat Bias
Day 2	Threat Bias	Threat Bias	Threat Bias
Day 3	Threat Bias	Threat Bias	Threat Bias

Screenshot of the QPF forecast verification for the Hurricane Sandy of the Weather Prediction Centre verification webpage (National Weather Service, NOAA: http://www.wpc.ncep.noaa.gov/tropical/case_studies/sandy_2012/sandyprecip.php):



Hurricane Sandy Precipitation Graphics

Forecast and Observed Precipitation for the 192 hours from 12Z October 24 through 12Z November 1, 2012

WPC QPF
(accumulation of seven 12Z issuances of the Day 1 24-hour QPF)

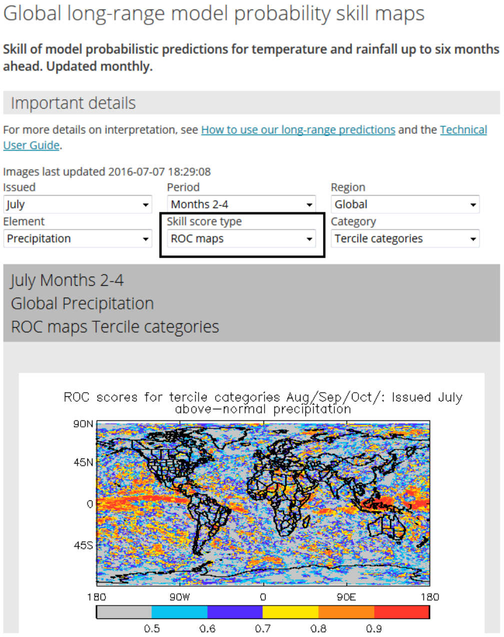
HPC 192-Hour Total QPF (from Day 1 forecasts)
VALID: 12Z OCTOBER 24, 2012 - 12Z NOVEMBER 1, 2012

STAGE IV QPE

STAGEIV 192-Hour Total QPE
VALID: 12Z OCTOBER 24, 2012 - 12Z NOVEMBER 1, 2012

IMPREX has received funding from the European Union Horizon 2020 Research and Innovation Programme under Grant agreement N° 641811

Screenshots from the MetOffice webpage on Global long-range model probability skill (http://www.metoffice.gov.uk/research/climate/seasonal-to-decadal/gpc-outlooks/glob-seas-prob-skill):



Global long-range model probability skill maps

Skill of model probabilistic predictions for temperature and rainfall up to six months ahead. Updated monthly.

Important details

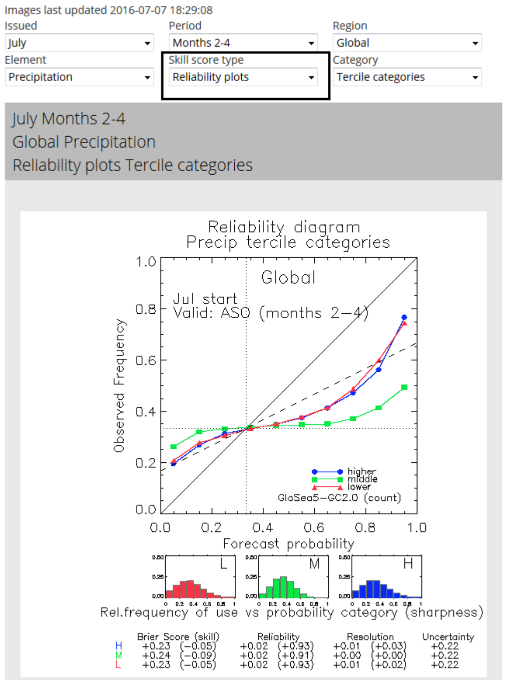
For more details on interpretation, see How to use our long-range predictions and the Technical User Guide.

Images last updated 2016-07-07 18:29:08

Issued: July
Period: Months 2-4
Region: Global
Element: Precipitation
Skill score type: ROC maps
Category: Tercile categories

July Months 2-4
Global Precipitation
ROC maps Tercile categories

ROC scores for tercile categories Aug/Sep/Oct/: Issued July
above−normal precipitation
90N
45N
0
45S
180
90W
0
90E
180
0.5
0.6
0.7
0.8
0.9



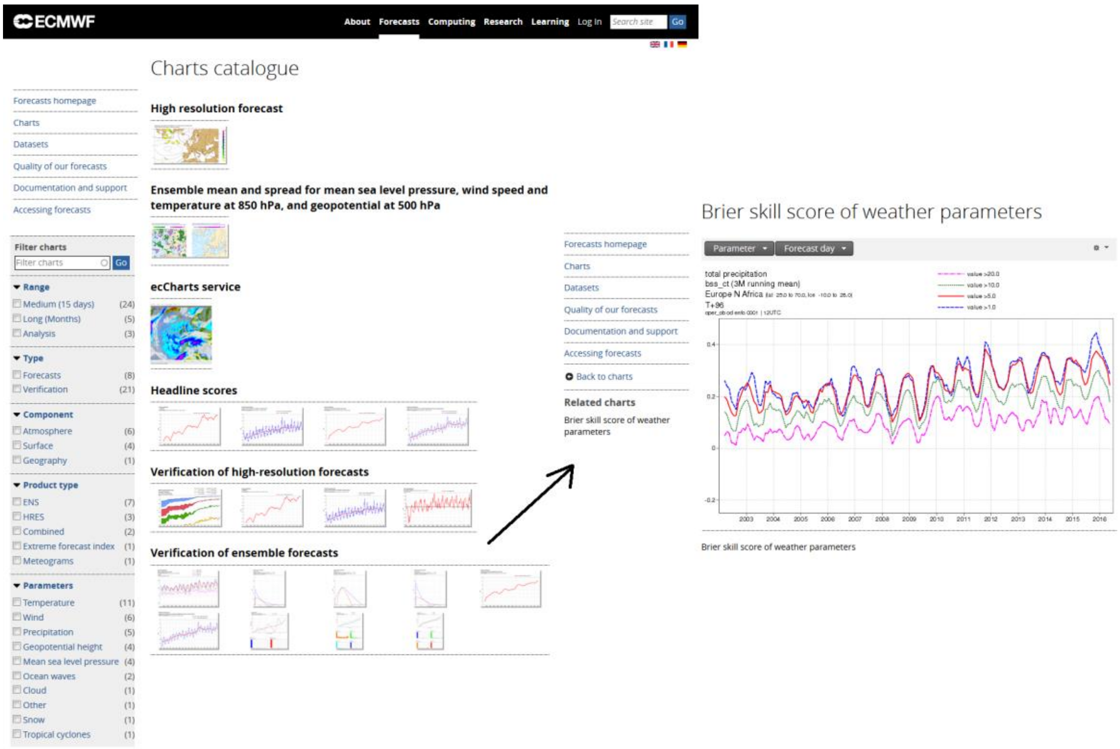
Images last updated 2016-07-07 18:29:08

Issued: July
Period: Months 2-4
Region: Global
Element: Precipitation
Skill score type: Reliability plots
Category: Tercile categories

July Months 2-4
Global Precipitation
Reliability plots Tercile categories

	Brier Score (skill)	Reliability	Resolution	Uncertainty
H	+0.23 (−0.05)	+0.02 (+0.93)	+0.01 (+0.03)	+0.22
M	+0.24 (−0.09)	+0.02 (+0.91)	+0.00 (+0.00)	+0.22
L	+0.23 (−0.05)	+0.02 (+0.93)	+0.01 (+0.02)	+0.22

Screenshots from the ECMWF website (http://www.ecmwf.int/en/forecasts/charts/catalogue):



Charts catalogue

Filter charts
Filter charts Go

▼ Range
☐ Medium (15 days) (24)
☐ Long (Months) (5)
☐ Analysis (3)

▼ Type
☐ Forecasts (8)
☐ Verification (21)

▼ Component
☐ Atmosphere (6)
☐ Surface (4)
☐ Geography (1)

▼ Product type
☐ ENS (7)
☐ HRES (3)
☐ Combined (2)
☐ Extreme forecast index (1)
☐ Meteograms (1)

▼ Parameters
☐ Temperature (11)
☐ Wind (6)
☐ Precipitation (5)
☐ Geopotential height (4)
☐ Mean sea level pressure (4)
☐ Ocean waves (2)
☐ Cloud (1)
☐ Other (1)
☐ Snow (1)
☐ Tropical cyclones (1)

High resolution forecast

Ensemble mean and spread for mean sea level pressure, wind speed and temperature at 850 hPa, and geopotential at 500 hPa

ecCharts service

Headline scores

Verification of high-resolution forecasts

Verification of ensemble forecasts

Brier skill score of weather parameters

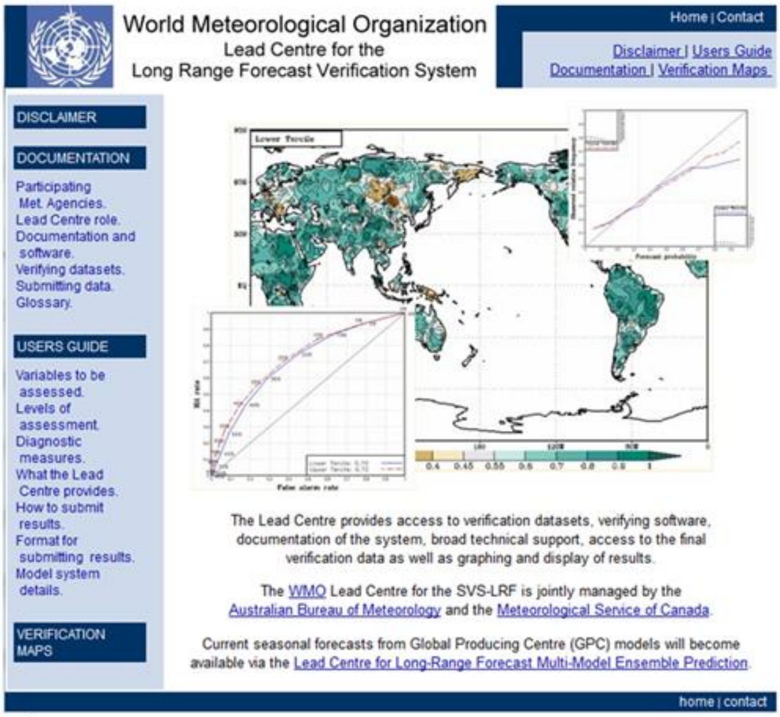
Related charts

Brier skill score of weather parameters

total precipitation
bss_ct (3M running mean)
Europe N Africa (lat 25.0 to 70.0, lon -10.0 to 28.0)
T+96
oper_ob od enfo 0001 | 12UTC

Brier skill score of weather parameters

Screenshots from the WMO website on its Long-range forecast verification system jointly managed by the Australian Bureau of Meteorology and the Meteorological Service of Canada (http://www.bom.gov.au/wmo/lrfvs/):



World Meteorological Organization
Lead Centre for the
Long Range Forecast Verification System

Home | Contact

Disclaimer | Users Guide
Documentation | Verification Maps

DISCLAIMER

DOCUMENTATION

Participating Met. Agencies.
Lead Centre role.
Documentation and software.
Verifying datasets.
Submitting data.
Glossary.

USERS GUIDE

Variables to be assessed.
Levels of assessment.
Diagnostic measures.
What the Lead Centre provides.
How to submit results.
Format for submitting results.
Model system details.

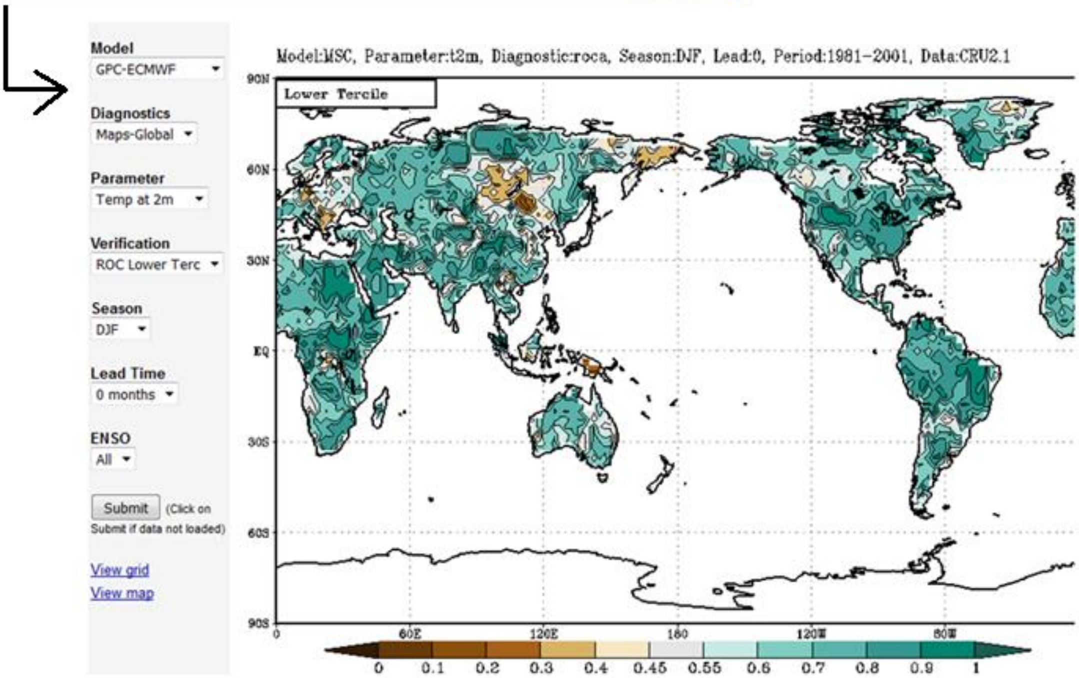
VERIFICATION MAPS

The Lead Centre provides access to verification datasets, verifying software, documentation of the system, broad technical support, access to the final verification data as well as graphing and display of results.

The WMO Lead Centre for the SVS-LRF is jointly managed by the Australian Bureau of Meteorology and the Meteorological Service of Canada.

Current seasonal forecasts from Global Producing Centre (GPC) models will become available via the Lead Centre for Long-Range Forecast Multi-Model Ensemble Prediction.

home | contact



Model: GPC-ECMWF

Diagnostics: Maps-Global

Parameter: Temp at 2m

Verification: ROC Lower Terc

Season: DJF

Lead Time: 0 months

ENSO: All

Submit (Click on Submit if data not loaded)

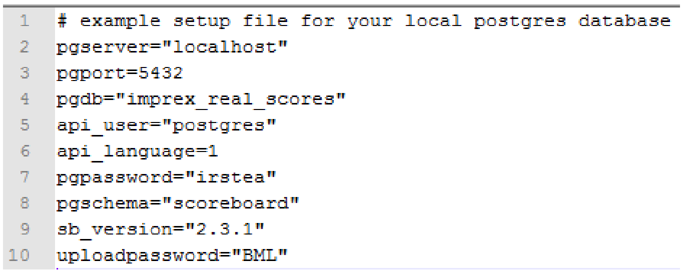
View grid
View map

Model:MSC, Parameter:t2m, Diagnostic:roca, Season:DJF, Lead:0, Period:1981-2001, Data:CRU2.1

IMPREX has received funding from the European Union Horizon 2020 Research and Innovation Programme under Grant agreement N° 641811

6.3 Annex 3 – Example of *.Renviron* file

This file should be located in the ~/R home directory of RStudio, or the file location must be updated in server.R and ui.R.



```
# example setup file for your local postgres database
pgserver="localhost"
pgport=5432
pgdb="imprex_real_scores"
api_user="postgres"
api_language=1
pgpassword="irstea"
pgschema="scoreboard"
sb_version="2.3.1"
uploadpassword="BML"
```

6.4 Annex 4 – Format specifications for input files to the IMPREX score database

Below is the file specification with individual descriptions (available in a Read_Me directory).

Files formats and needed data and metadata

The scoreboard works with RDS and text files.

One single data file can only contain data from one experiment evaluated on one score, with one initialisation, but it can include several stations.

NA values must be entered as either:

- NA
- -9999.

I invite you to open these files to better know what to put in. I provide here some important elements of explanation:

- The RDS files are R data files. They contain a list of elements containing the metadata and data you want to upload.
- The text file contains several lines with semi-column separated metadata (8 fields needed) and semi-column separated data.
- In both formats, the metadata are read as characters.
  - Provider : your name
  - CaseStudy: your case study (“Jucar River Basin” for instance)

- ForecastSystem: your forecast system
- ForecastType: the name you want to give to your experiment, e.g. the name of your model, with its time step if you can have several time steps. It is important to give a different name if you want to compare the new configuration with an older one. The score name should not appear here.
- ModelVariable: the name of the variable on which the score is evaluated.
- ScoreType: the name of the score
- ScoreLeadTimeUnit: "Day", "Week", "Month", etc. are examples. If every 15 days you make a seasonal forecast at a monthly time step for the next 6 months, you must put "Month".
- VerificationPeriod: on which period you computed your score
- Data:
  - LocationID: whatever code you use to design your station / grid point / area on which the score is computed. This is read as character
  - River_names: read as characters
  - Station_names: read as characters
  - Latitude: read as numeric
  - Longitude: read as numeric
  - Score_Value: your actual score values! Read as numeric
  - LeadTime: the leadtime that is evaluated. If every 15 days you make a seasonal forecast at a monthly time step for the next 6 months, you must put here values from 1 to 6.

- Important notice: some fields must be very carefully filled. Indeed, the database uses some of them to link the different tables of the database. The fields are all the metadata fields (except VerificationPeriod for now) and the LocationID. To illustrate what I am saying, if provider 1 gives CRPSS and provider 2 provides CRPSkillScore, which are basically the same scores, they will not be comparable in the 3rd panel of the scoreboard.

IMPREX has received funding from the European Union Horizon 2020 Research and Innovation Programme under Grant Agreement N° 641811