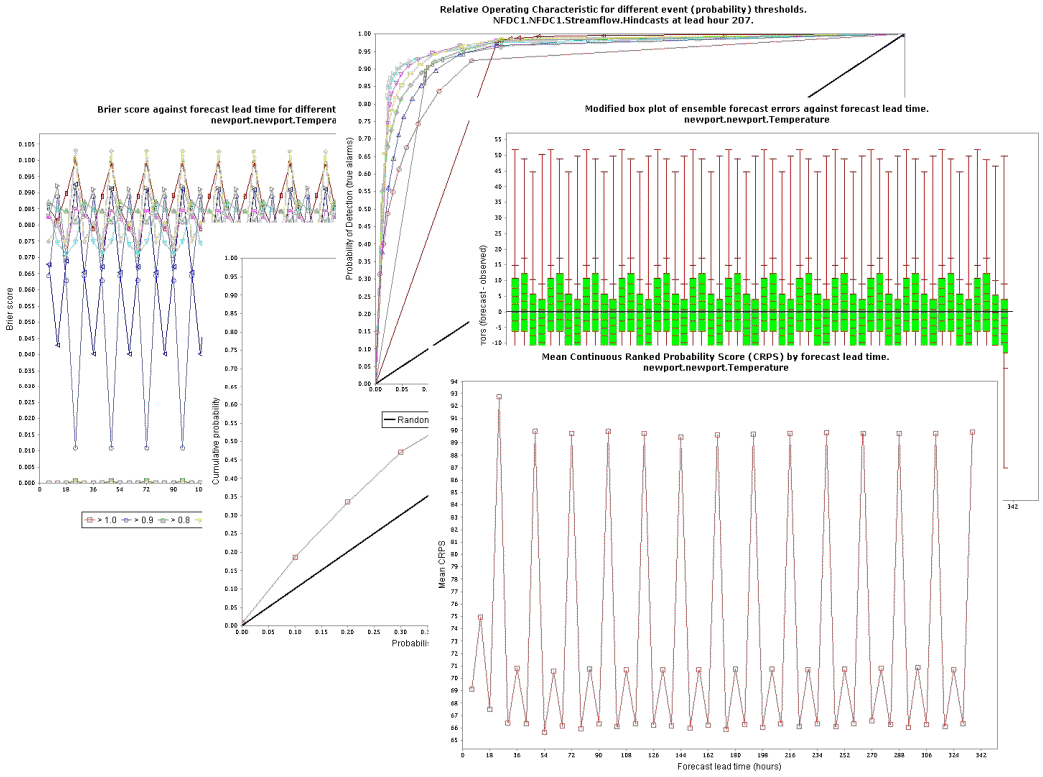
Ensemble Verification System (EVS)

Version 1.0



User's Manual

James D. Brown, Julie Demargne, Limin Wu, Dong-Jun Seo

Hydrologic Ensemble Prediction Group, Office of Hydrologic Development, National Weather Service, National Oceanic and Atmospheric Administration, 1325 East-West Highway, Silver Spring, Maryland, 20910, USA; e-mail: James.D.Brown@noaa.gov

Preface

The Ensemble Verification System (EVS) is an experimental prototype developed at OHD for verifying ensemble forecasts of hydrologic and hydrometeorological variables, such as temperature, precipitation, streamflow and stage. It is intended for use by forecasters at the River Forecast Centers (RFC), researchers and developers at OHD, and collaborators elsewhere. EVS is intended to be flexible, modular and open to accommodate enhancements and additions not only by its developers but also by its users. As such, in addition to comments and suggestions for improvement by the EVS development team at OHD, we welcome participation of the users in the continuing development of EVS toward a versatile and easy-to-use operational ensemble verification capability.

The Hydrologic Ensemble Prediction Group (HEP)
Dong-Jun Seo, Group Leader
Julie Demargne[2]
Limin Wu
James Brown[1]
Haksu Lee
Satish Regonda

[1] EVS Primary Point of Contact, James.D.Brown@noaa.gov, 301-713-0640 ext 224
[2] EVS Secondary Point of Contact, Julie.Demargne@noaa.gov, 301-713-0640 ext 162

Acknowledgments

This work was supported by the NOAA's Advanced Hydrologic Prediction Service (AHPS) and Climate Prediction Program for the Americas (CPPA).

Contents

1. INTRODUCTION

The National Oceanic and Atmospheric Administration's (NOAA's) National Weather Service (NWS) requires systematic verification of hydrologic and hydrometeorological ensemble forecasts. This is necessary to generate reliable and skilful products. Such verification and validation will help forecasters estimate the quality of probabilistic forecasts according to lead times, forecast locations, and the ensemble prediction systems in use. They will also help assess the usefulness of ensemble products for end-users.

The Ensemble Verification System (EVS) aims to support the verification of ensemble forecasts and hindcasts of hydrometeorological variables (e.g. precipitation and temperature) and hydrological variables (e.g. streamflow). EVS is free software and has been developed in a modular framework to allow enhancements and additions by scientists, forecasters, and other users.

2. INSTALLATION AND START-UP

2.1 Requirements

In order to run EVS you will need:

1. The Java™ Runtime Environment (JRE) version 5.0 (1.5.0_12) or higher. You can check your current version of Java by opening a command prompt and typing java –version. If the command is not recognized, you do not have a version of the JRE installed. Otherwise, if it is older than 1.5.0_12, you should update the JRE. The JRE is free software and may be downloaded from the Sun website:

   http://java.sun.com/javase/downloads/index.jsp

2. The EVS executable, EVS.jar, and associated resources in EVS_1.0.zip;

3. Microsoft Windows 98/2000/NT/XP/Vista Operating System (OS) or Linux. In addition, you will need:

   - A minimum of 32MB of RAM and ~50MB of hard-disk space free.
   - For many practical applications of EVS, involving verification of large datasets more RAM may be required. A minimum of 512MB is recommended.

2.2 Unpacking and running EVS

Once you have obtained the EVS software, unpack the zipped archive to any directory of your computer (e.g. `C:/Program Files/EVS_1.0/`) using, for example, WinZip™ on Windows. Do not move the EVS.jar executable from the existing directory structure: create a shortcut elsewhere if required.

There are two possible ways of running EVS, namely: 1) by opening the Graphical User Interface (GUI); and 2) by executing EVS from the command line with a pre-defined project.

Executing EVS with the GUI:

Once you have unpacked the software, you may run EVS by double-clicking on "EVS.jar" in Windows or by navigating to the root directory and typing "java –jar EVS.jar" at a command prompt.

Executing EVS without the GUI:

In order to execute EVS without the GUI, you must have one or more pre-defined projects with valid verification units (and possibly aggregation units) stored inside them. EVS projects are defined in XML (see Appendix A2) and may be created with or without the GUI. For example, a base project may be created with the GUI and then perturbed with a script outside of the GUI (e.g. changing the verification unit name, input and output data sources). Each perturbation can then be executed from a script without invoking the GUI. One or more projects may be invoked from a command prompt by typing:

```
java –jar EVS.jar project_1.evs
```

where `project_1.evs` is an EVS project (the project need not be located in the root directory, but should be referenced by its full path otherwise). The list may be extended by adding projects with a space between each project. By default, the graphical and numerical results are written to the output directories specified in the projects. The numerical results are written in XML format and the graphical results are written as jpeg images.

2.3 Troubleshooting the installation

List of typical problems and actions:

- **"Nothing happens when executing EVS.jar"**

Ensure that the Java Runtime Environment (JRE) is installed on your machine and is in your PATH. The JRE should be version 5.0 (1.5.0_12) or higher. To check that a suitable version of the JRE is installed and in your PATH, open a command prompt and type:

java -version

If the command is not recognised, the JRE is not installed and in your PATH. If the version is below 5.0 (1.5.0_12) update the JRE (see above).

If this does not help, check the root directory of your installation for a log file named "evs.log". If the first line of the log file is:

com/incors/plaf/alloy/AlloyLookAndFeel

then EVS has been unable to load the resources required for proper execution of the software. Check that "EVS.jar" has not been moved from the original installation directory (i.e. that the internal structure of the archive "EVS_1.0.zip" is preserved).

Otherwise, send the error message to the authors for advice on how to proceed (James.D.Brown@noaa.gov).

– **"An error message is thrown when executing EVS.jar"**

If an error message is thrown by the JRE (i.e. a java error appears in the message), the error may be caused by the local installation of Java.

2.4 Altering memory settings

By default, the amount of RAM memory available to EVS is restricted by the Java Virtual Machine. In order to perform ensemble verification with large datasets, it may be necessary to override this default and increase the amount of memory available. This is achieved by executing EVS on the command line. Navigate to the installation directory of EVS, and type:

start javaw -jar -Xms**64**m -Xmx**500**m EVS.jar

where **64** (MB) is the minimum memory allocation *in this example* and **500** is the maximum allocation. The maximum memory allocation should be significantly lower than the total amount of RAM available on your machine, as other programs, including the operating system, will require memory to run efficiently.

2.5 Source code and documentation

The Java source code for EVS can be found in the "src.zip" archive in the root directory of your installation. The Application Programming Interface (API) is described in the html documentation, which accompanies the software (`/docs` directory).

3. OVERVIEW OF FUNCTIONALITY

3.1 Summary of functionality in EVS Version 1.0

The functionality currently supported by EVS includes:

- Pairing of observations (given in the observed file in datacard format) and ensemble forecast values (given in the ensemble files in datacard format or CS binary format) to perform verification for a given point; the observed and forecast values may be in different time systems, the time offset between the two systems being defined by the user;

- Computation of multiple verification statistics for different variables (precipitation, temperature, or streamflow) at a single point. The statistics may be computed for any number of lead days and forecast resolutions. The statistics currently include:
  - For deterministic verification using the ensemble mean: correlation coefficient, mean error, and root mean squared error
  - For probabilistic verification: Brier Score (BS); Continuous Ranked Probability Score (CRPS); Mean Capture Rate diagram (MCR); Modified box plots; Relative Operating Characteristic (ROC); Reliability diagram; and Talagrand diagram (rank histogram).

- Conditional verification based on: 1) a restricted set of dates (e.g. months, days, weeks, or some combination of these); 2) a restricted set of observed or forecast values (e.g. ensemble mean exceeding some threshold, maximum observed values within a 90 day window). Values may be specified in units of (observed) probability or in real values, such as flood stage.

- Pooling or 'aggregation' of observed-forecast pairs from a group of points with common verification parameters; the aggregate statistics are computed from the pooled pairs;

- Generation of graphics and numerical products, which may be written to file in various formats (e.g. jpeg files), plotted within EVS or both.

3.2 Planned functionality

The additional functionalities planned for future versions of EVS includes, in no particular order:

- Compute measures of uncertainty, such as confidence intervals, for the verification statistics. In order to present these in meaningful ways, it should be possible to answer questions such as 'Can I apply conditions X and Y to my verification dataset, and still have confidence Z in the results?' Here, conditions X and Y might involve the selection of forecasts where flow exceeds a given level, or for winter months only;

- Additional options for generating verification products, such as plots.

- Functionality for verifying joint distributions; that is, maintaining the relationships between points in space and time (e.g. to verify errors in the timing of a discharge event or in the reproduction of spatial patterns).

- Ability to compute metrics for arbitrary reference forecasts, such as climatology, persistence or raw model output (e.g. before data assimilation or manual adjustment), and derive measures of skill based on arbitrary skill functions (e.g. ratio of one metric over another).

- Development of a batch language to support generation of verification products without running the GUI. For example, it should be possible to create a template point and apply this to a wider group of forecast points, changing only the observed and forecast data sources via a batch processor.

- To fully integrate EVS within the Experimental Ensemble Forecasting System (XEFS), which is currently being developed at NOAA. The XEFS will comprise a coordinate suite of tools for end-to-end probabilistic forecasting of hydrological and hydro-meteorological variables. Capabilities will include the generation of model inputs, updating of model state variables and parameters, verification of outputs, and generation of ensemble products for end-users.

4. GETTING STARTED

4.1 Performing ensemble verification with EVS

Performing ensemble verification with EVS is separated into three stages, namely:

1. VERIFICATION: Defining one or more 'verification units', where each unit currently comprises a time series of a single variable at one point (e.g. a river segment), together with the verification statistics to compute;
2. AGGREGATION: Defining one or more 'aggregation units', where each unit comprises one or more verification units whose data will be pooled;
3. OUTPUT: Generation of products, such as numerical results and plots of statistics from verification and/or aggregation;

These stages are separated into 'panels' in the user interface. To begin with, a verification study with EVS may involve linearly navigating through these panels using the "**Next**" and "**Back**" buttons. After one or more verification or aggregation units have been defined and saved, the route of entry into the software may vary. For example, it might involve modifying and saving an existing unit for later use or generating new plots of existing statistics.

4.2 Administrative functions

The opening window of EVS, together with the Taskbar, is shown in *figure 1*. The opening window displays the verification units loaded into the software. The Taskbar is visible throughout the operation of EVS and is used for administrative tasks, such as creating, opening, closing and saving a project. The Taskbar options are listed in *table 1*. Shortcuts are provided on the Taskbar for some common operations, but all operations are otherwise accessible through the dropdown lists.

Figure 1: The opening window of EVS

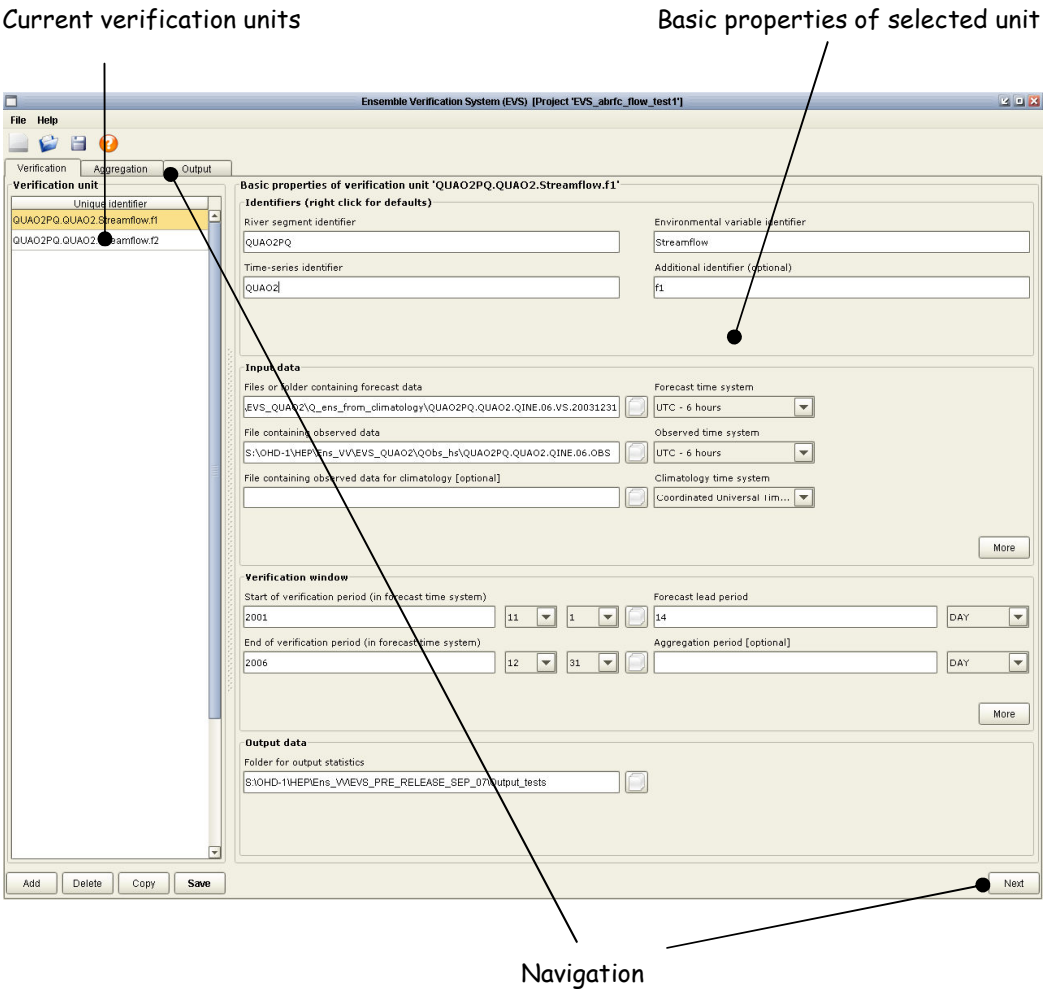
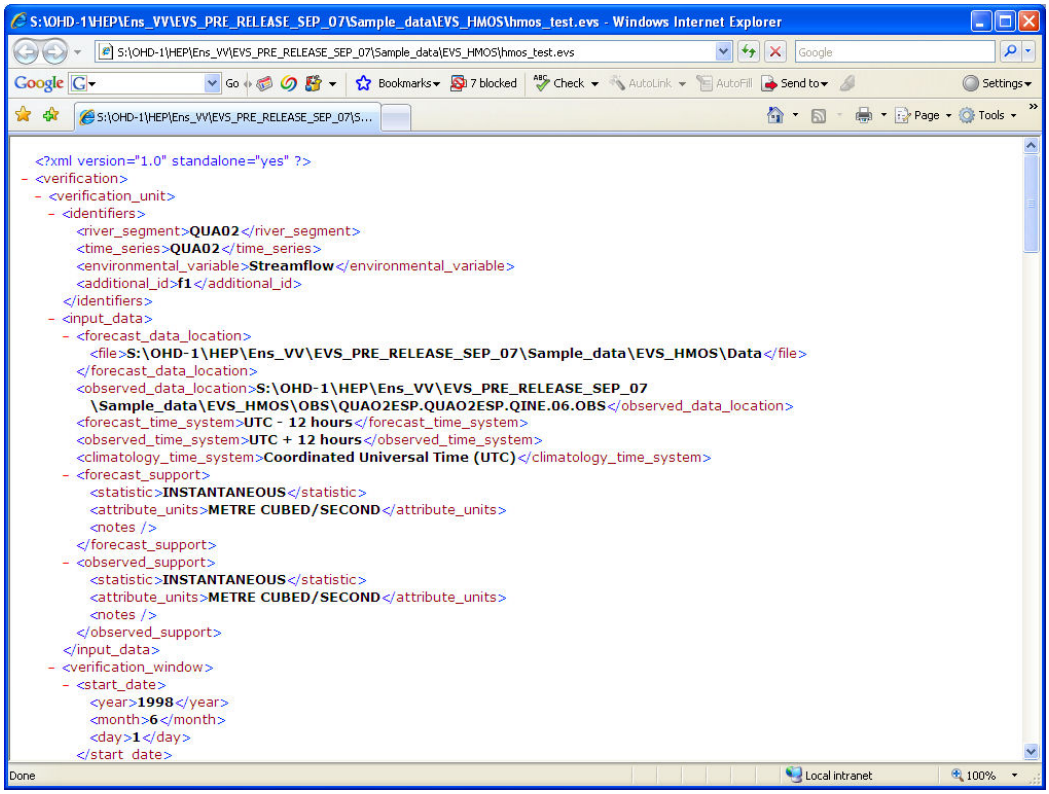


Table 1: Menu items

Menu	Function	Use
File	New project	Creates a new project
	Open project	Opens a project file (*.evs)
	Close project	Closes a project
	Save project	Updates or creates a project file (*.evs)
	Save project as	Updates or creates a named project file (*.evs)
	Exit	Exits EVS
Help	Messages on/off	Displays/hides tool tips
	Console	Shows the details of errors thrown
	About	Credits

4.3 Creating projects

All work within EVS (including user interface settings) can be saved to a project file with the .evs extension. A new project is created with the **New project** option under the **File** dialog. An existing project is saved using the **Save** or **Save As**… options. These options are also available on the Taskbar. Project files are stored in XML format and may be opened in an Internet browser or text editor. An example is given below:



4.4 A step-by-step guide to the windows in EVS

<u>Verification: window 1</u>

The first stage of ensemble verification requires one or more 'verification units' to be defined (*figure 1*). In this context, a verification unit (VU) comprises a time-series of a single variable at one point, hereafter assumed to be a river segment. A VU is uniquely identified by these three attributes, which must be entered in the first window, and are then displayed in the table and identifiers panel. A new VU may be added to the current project by clicking "**Add**" in the bottom left corner of the window (*figure 1*). This adds a VU with some default values for the identifiers. On entering multiple VUs, the basic properties of the *selected* VU (i.e. the item highlighted in the table) will be shown in the panels on the right. Existing units may be deleted or

copied by selecting an existing unit in the table and clicking "**delete**" or "**copy**", respectively. On copying a unit, all of the properties of the unit are copied *except* the identifiers, which must be unique. This provides a convenient way to specify multiple units with the same verification properties (multiple segments to be verified for the same variable with the same temporal parameters).

The VU is defined by four different dialogs: Identifiers, Input data, Verification window, and Output data.

Identifiers dialog:

- River segment identifier: segment ID (referred to as segment_id);
- Time series identifier: time series ID (referred to as time_series_id);
- Environmental variable identifier: e.g. precipitation, temperature, streamflow;
- Additional identifier: allows, for example, identification of a forecast system;

The names of the river segment and time-series are unrestricted (aside from a blank name or a name containing the illegal character '.' used to separate the identifiers). The default names for the environmental variable are provided by right-clicking on the variable identifier box.

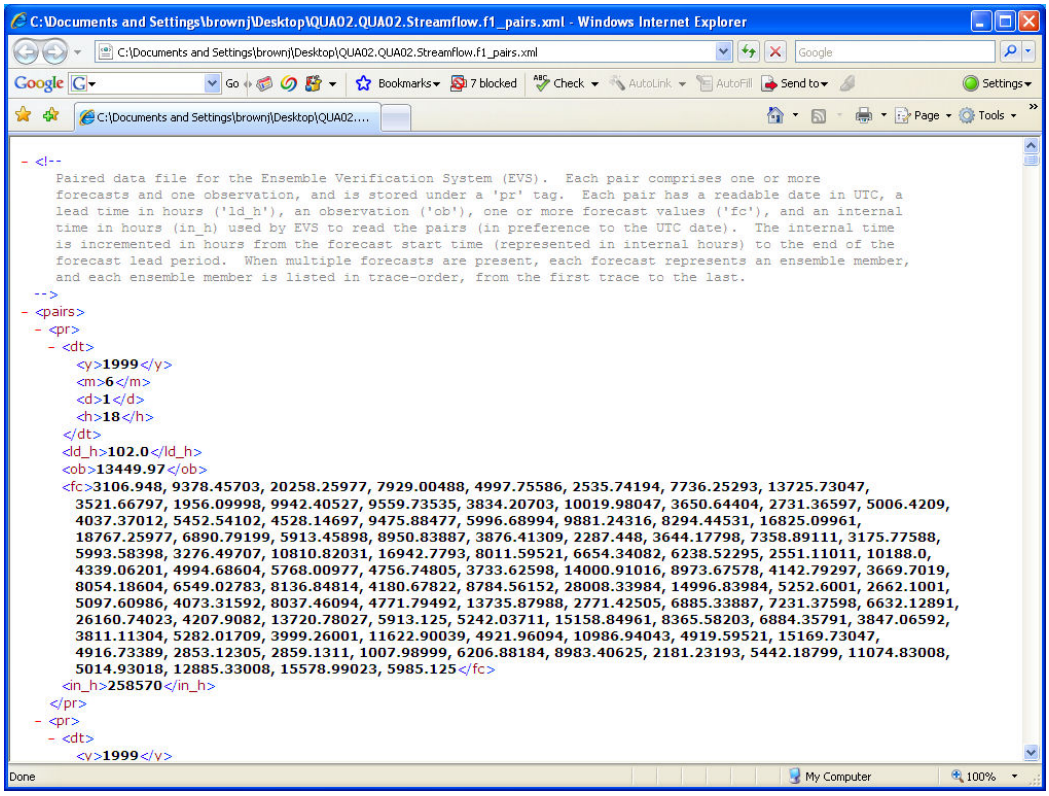
Input data dialog:

- Files or folder containing forecast data: path to the folder containing the ensemble forecasts (all files will be read from this directory), or a file array chosen through the associated file dialog;
- File containing observed data: path to concurrent observations of that variable, which are used to verify the forecasts;
- File containing observed data for climatology: option to load a separate file from which to determine climatology;
- Time systems: the time systems for the observations, forecasts and climatological data. The time systems of the forecasts and observations are required for pairing these data (on the basis of time);

The paths may be entered manually or by clicking on the adjacent button, which opens a file dialog.

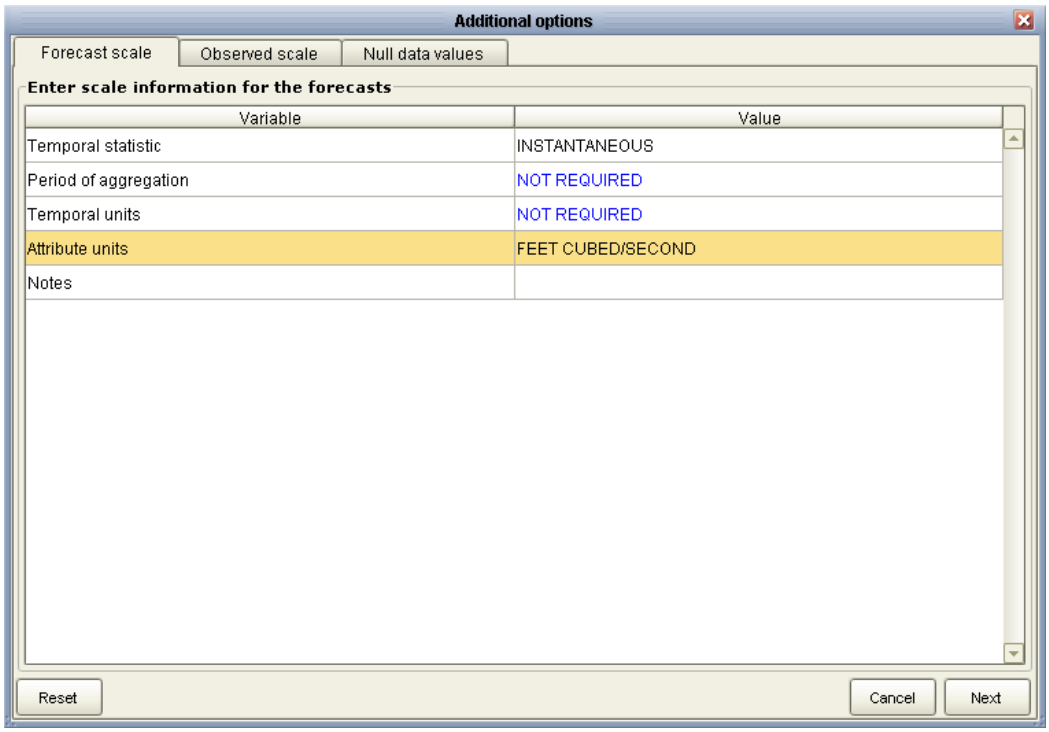
When conducting verification for the first time, the observations and forecasts are paired. These pairs are used to compute the differences between the observed and forecast values (i.e. the forecast 'errors') at concurrent times. For subsequent work with the same unit, no pairing is necessary unless some of the input parameters have changed (e.g. the verification window). The paired data are stored in XML format,

which may be opened in an Internet browser or text editor. Each forecast-observation pair is stored with a date in UTC (year, month, day, and hour), the forecast lead time in hours, the observation, and the corresponding forecast ensemble members. A detailed explanation is also provided in the paired file header. An example of a paired file is given below:



Throughout EVS, default options are presented immediately to the user. In some cases, additional ('advanced') options are presented in sub-windows, accessible through the main windows. For example, the "**More**" button within the Input data dialog opens a window for entering information about the scales at which the forecasts and observations are defined, among other things. Scale information includes the units of measurement (e.g. cubic feet/second) and temporal support at which the forecasts and observations are recorded (e.g. instantaneous vs. time-averaged). The forecasts and observations must be defined at equivalent scales for a meaningful comparison between them. By default, the scales are assumed to be equivalent. However, in the absence of user-defined information, a warning message will be presented on conducting verification. This warning message is avoided if the scale information is entered explicitly. An example of the 'Additional options' dialog is given in *figure 2*. In addition to the scale of the forecasts and observations, the identifier for 'null' or missing values can be changed (i.e. values ignored during all processing, including metric calculation). By default, the null value is -999.

Figure 2: The Additional options dialog, accessed from the first verification window



Verification window:

- Start of verification period (in forecast time system): the start date for verification purposes. This may occur before or after the period for which data are available. Missing periods will be ignored. The verification period should respect the forecast time system, in case it differs from the observed time system. The start date may be entered manually or via a calendar utility accessed through the adjacent button;
- End of verification period (in forecast time system): as above, but defines the last date to consider;
- Forecast lead period: at each forecast time, a prediction is made for a period into the future. This period is referred to as the forecast lead period. For example, if the forecasts are issued every 6 hours and extend 14 days into the future, the forecast lead period is 14 days. The forecast lead period may be used to narrow the range of forecasts considered (e.g. only the first 5 lead days when the forecasts are available for 90 lead days).
- Aggregation period: when evaluating long-term ensembles (e.g. with a 1 year lead period), verification results may be confused by short-term variability, which is not relevant for the types of decisions that inform long-term forecasts, such as water availability in a reservoir. Aggregation of the forecasts and observations allows short-term variability to be removed by

averaging over the period that *does* matter for decision making purposes. For example, daily forecasts may be aggregated into 90-day averages (assuming the forecast lead period is at least 90 days).

The verification window may be refined using conditions on the dates and sizes of the observed or forecast values considered. These 'advanced' options are accessed through the "**More**" button in the Verification window. For example, verification may be restricted to 'winter months' within the overall verification period, or may be limited to forecasts whose ensemble mean is below a given threshold (e.g. zero degrees for temperature forecasts). When conditioning on variable value, conditions may be built for the current unit (selected in the main verification window) using the values of another variable (e.g. select streamflow when precipitation is non-zero), providing the variables have the same prediction dates and intervals. Such conditioning may be relatively simple or arbitrarily complex depending on how many conditions are imposed simultaneously. However, there is a trade-off between the specificity of a verification study, which is increased by conditioning, and the number of samples available to compute the verification statistics, which is reduced by conditioning (i.e. sampling uncertainty is increased). The dialog for conditioning on date and variable value is shown in *figures 3a* and *3b*, respectively.

Figure 3a: Dialog for refining verification window: conditioning with dates

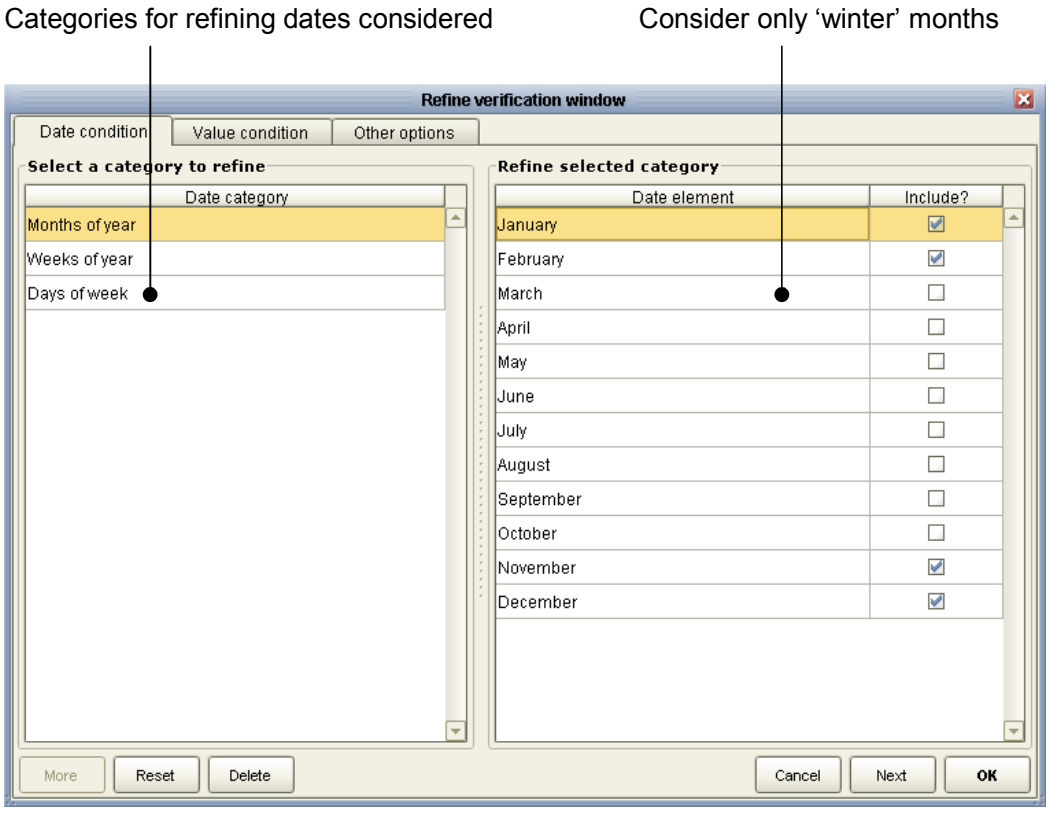


Figure 3b: Dialog for refining verification window: conditioning with variable value

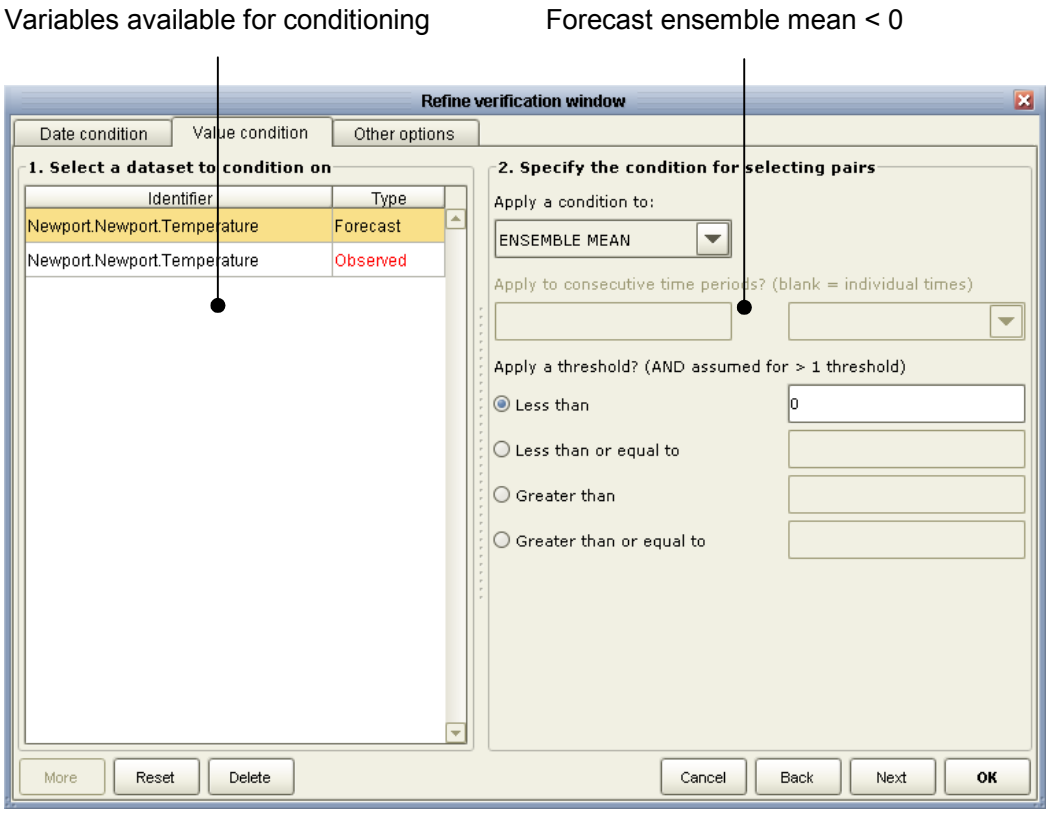
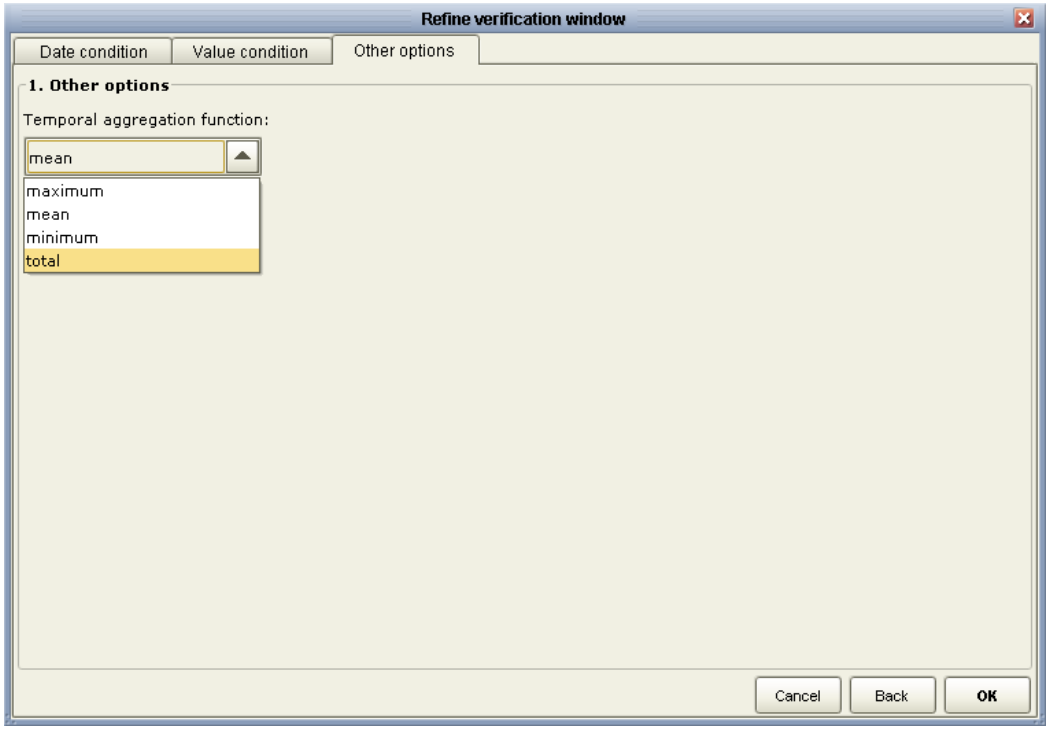


Figure 3c: Dialog for refining verification window: other options



Additional refinement options are available in the "Other options" section of the refinement dialog. Currently, there is only one additional option, which allows the temporal aggregation function to be defined. By default, aggregations requested in the main verification window involve a mean average over the specified period. This may be changed to a total (i.e. accumulation), minimum or maximum value (*figure 3c*).

Output data:

- Folder for output statistics: path to the folder for writing the paired files and the verification output data generated by the system, if written output is requested (see below).

Verification: window 2

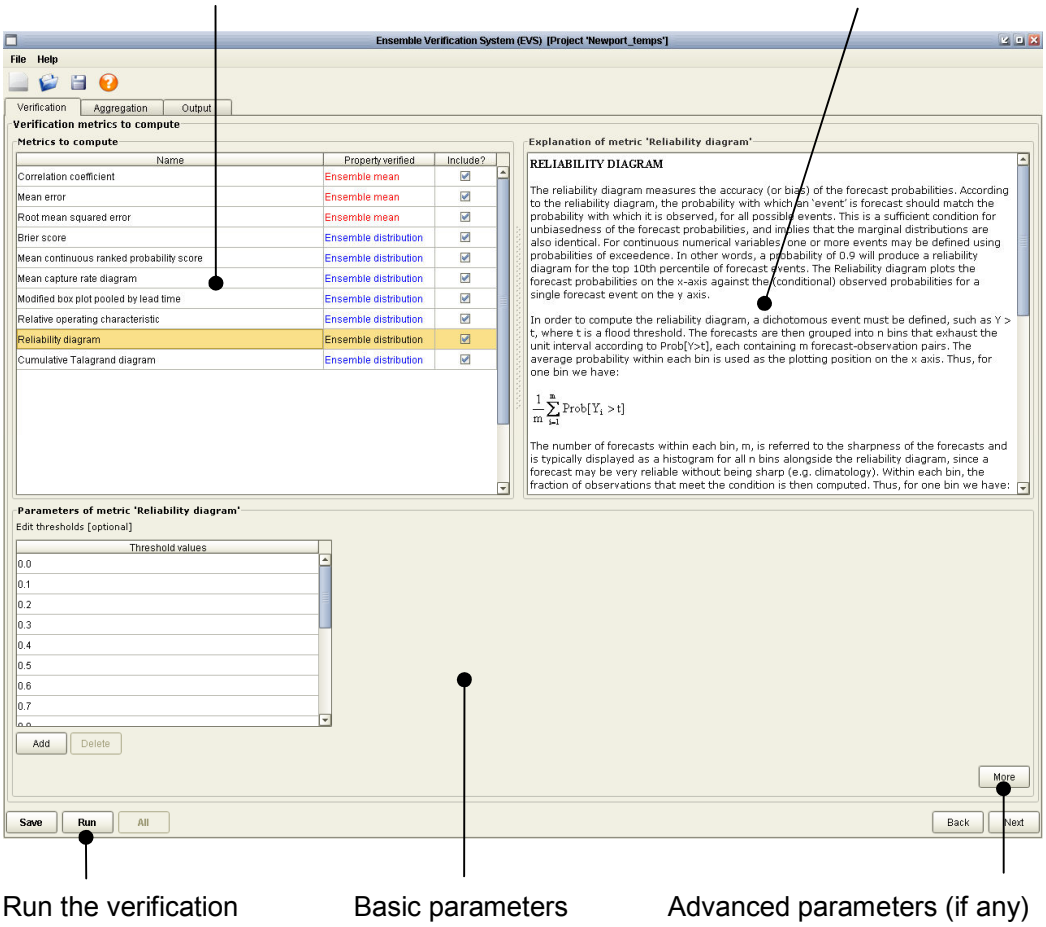
The second window in the Verification pane (*figure 4*) is accessed by clicking "**Next**" from the first window (*figure 1*).

On selecting a given statistic in the table, information about that statistic is provided in the top right dialog, and the parameters of the statistic are displayed for entering/editing in the bottom-left panel. A statistic is included, and its parameter values are enabled for editing, by checking the box adjacent to the statistic in the top left table. The parameters of each metric are listed in *table 2*.

Many of the ensemble statistics have thresholds, which are used to compare the observed values and the forecast values. These thresholds define a subset of data from which the metric is calculated. Most of the metrics can be computed from *all* data, as well as subsets of data defined by the thresholds. However, some metrics, such as the reliability diagram, relative operating characteristic and Brier score, *require* one or more thresholds to be defined, and cannot be computed from all data. For these metrics, the thresholds represent cutoff values within the observed probability distribution from which discrete events are computed. By default, the thresholds refer to exceedence probabilities (i.e. all data > than the threshold) within the observed probability distribution and must, therefore, vary between 0-1. For example, a threshold of 0.2 would select all data whose observations lie above a probability of 0.2 in the observed probability distribution. The thresholds may be modified by entering new values into the table or by deleting thresholds and adding new ones. The types of thresholds may be modified via the "**More**" button, which displays an advanced options dialog. For example, the thresholds may be changed to real-values, rather than probabilities (e.g. Flood Stage) and the logical condition can be changed to non-exceedence, among others (see below also).

After modifying the verification statistics and their parameters, the new information is saved to the current unit by clicking "**Save**".

Figure 4: The second window in the Verification pane



In addition to the metrics shown in the second window of the Verification pane (*figure 4*), other verification metrics may be available for research purposes via commands added to the evs project file (see *Appendix A1*). These metrics are *not* fully supported and are subject to change, including removal at the discretion of the software developers.

Depending on the selected verification metric, there may be some additional, advanced, parameters that can be altered. These parameters are available through the **"More"** button, which will become enabled if more parameters are available. For example, when computing ensemble metrics using probability thresholds, the thresholds may be treated as non-exceedence (<, <=) or exceedence probabilities (>, >=), which may be useful for exploring low- versus high-flow conditions, respectively

(*figure 5*). The parameter options for each metric are summarized in *table 2*. A 'basic' parameter is accessed through the main window in EVS, while an 'advanced' parameter is accessed through the "**More**" button (as in *figure 5*).

Figure 5: Advanced parameter options for a selected metric (reliability in this case)

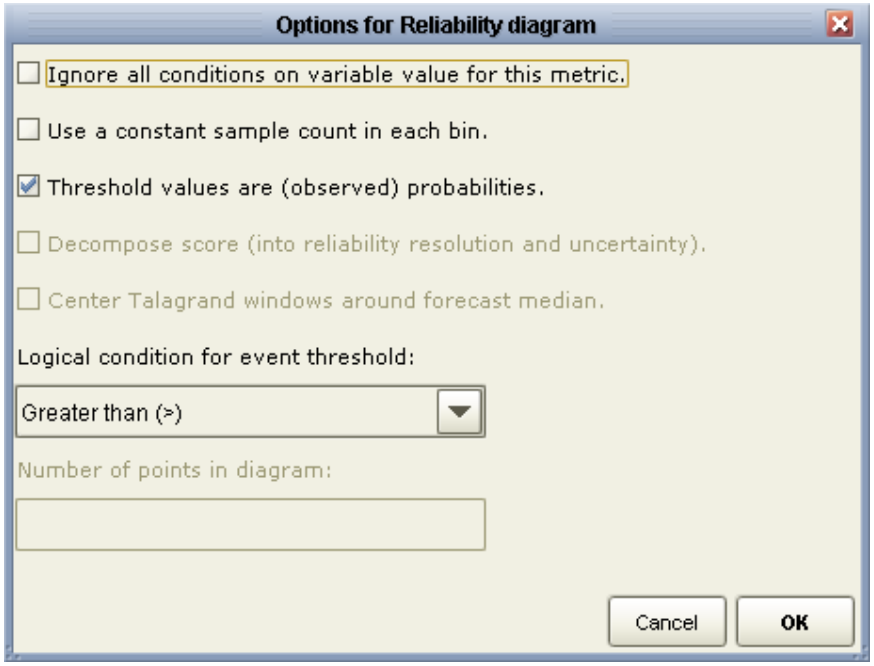


Table 2: Parameters for each verification metric

Metric	Parameter (and type)	Meaning
Mean error	Thresholds **(basic)**	Produces the metric for each subset of data specified by the threshold. Thresholds *always* refer to the observations. By default, they refer to exceedence probabilities from the observed distribution (i.e. climatology).
	Ignore conditions on variable value **(advanced)**	Any conditions on the observed or forecast values used to subset pairs (an advanced option in the first Verification window) will be ignored for this metric.

	Threshold values are observed probabilities **(advanced)**	If this parameter is true (checked; the default option), the threshold parameter (above) will refer to probabilities in the observed probability distribution. For example, a threshold value of 0.2 would select pairs in relation to the real value corresponding to probability 0.2 in the observed probability distribution. The form of the relationship will depend on the logical condition for the threshold (below). If this parameter is false (unchecked), the thresholds are interpreted as real-values in observed units (e.g. cubic feet per second).
	Logical condition for event threshold **(advanced)**	Changes the logical condition for any thresholds used to subset data. For example, if the logical condition is "greater than", only those forecast - observation pairs whose observed values are greater than the threshold will be used.
Root Mean Squared Error	Same as mean error.	Same as mean error.
Correlation coefficient	Same as mean error.	Same as mean error.
Brier score	Same as mean error.	Same as mean error.
Mean Continuous Ranked Probability Score	Same as mean error.	Same as mean error.
Mean Capture Rate Diagram.	Same as mean error.	Same as mean error.
	Number of points in diagram **(advanced)**	Sets the number of equally-spaced probability values (from 0-1) for which the metric will be computed and plotted.
Modified box plot pooled by lead time.	Ignore conditions on variable value **(advanced)**	Same as parameter for mean error.
	Number of points in diagram **(advanced)**.	Sets the number of equally-spaced probability values (from 0-1) at which the boxes will be computed and plotted. The middle thresholds form the boxes and outer thresholds form the whiskers.
Relative Operating Characteristic	Same as Mean Capture Rate Diagram.	Same as Mean Capture Rate Diagram.
Reliability Diagram	Ignore conditions on variable value **(advanced)**	Same as parameter for mean error.

	Use a constant sample count in each bin **(advanced)**.	If this parameter is false (unchecked; the default option), the forecasts probability bins for which the reliability values are computed will take a fixed width in the range 0-1 depending on the number of points requested for the diagram (below). If this parameter is true (checked), the forecast probability bins for which the reliability values are computed will vary in width such that each bin captures the same number of forecasts.
	Threshold values are observed probabilities **(advanced)**.	Same as parameter for mean error.
	Logical condition for event threshold **(advanced)**.	Same as parameter for mean error.
	Number of points in diagram **(advanced)**.	Sets the number of probability bins (from 0-1) for which the metric will be computed and plotted. These bins may capture an equal sample count (see above) or may be equally spaced.
Talagrand diagram	Ignore conditions on variable value **(advanced)**.	Same as parameter for mean error.
	Threshold values are observed probabilities **(advanced)**.	Same as parameter for mean error.
	Center Talagrand windows around forecast median **(advanced)**.	If this parameter is false (unchecked; the default option), the probability of an observation falling within a forecast bin is determined for bins separated by probabilities within the forecast distribution. For example, if the parameter for the 'Number of points in the diagram' (see below) is 10, probabilities will be determined for bins representing deciles of the forecast. If this parameter is true (checked), probabilities of the observation falling within a forecast bin will be determined for symmetric forecast bins defined with respect to the forecast median.
	Logical condition for event threshold **(advanced)**.	Same as parameter for mean error.
	Number of points in diagram **(advanced)**.	Defines the number of forecast bins for which the probability of an observation falling within that bin is determined.

All of the information necessary to verify the ensemble forecasts is now available, and the verification may be executed by clicking "**Run**" for the current segment or "**All**" to execute verification for all available segments. This may take several minutes or longer, depending on the size of the datasets involved. If not already available, the paired files are created (see above) and the selected metrics are then computed for each unit. No products are displayed or written at this stage; instead the numerical results are stored in memory, in preparation for generating these products (see below).

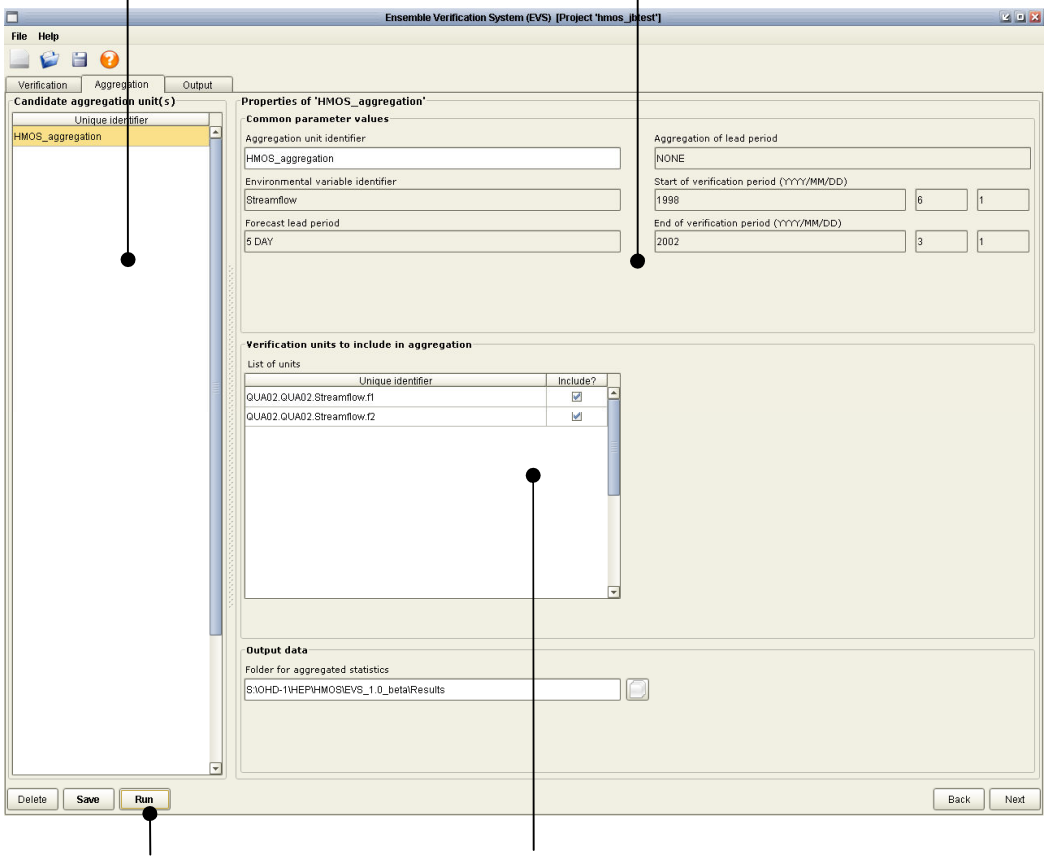
Aggregation: window 1

Alongside verification of ensemble forecasts from a single point, it is possible to aggregate verification statistics across multiple river segments. This is achieved in the first aggregation window (*figure 6*). Only those points for which aggregation is possible will be displayed in the aggregation window (i.e. with common parameter values). If no aggregation units (AUs) are displayed, no comparable VUs have been defined. The properties of an AU may be viewed or edited by selecting the unit in the table. Each AU is given a default identifier, which may be altered by the user. Multiple AUs may be defined in one project to generate aggregate statistics on various groups of river segments with common verification parameters. Aggregation is achieved by simply pooling the pairs from multiple VUs and applying the statistics to the pooled dataset.

Figure 6: The first window in the Aggregation pane

List of aggregation units

Common parameters of the VUs in the AU



Run the aggregation

Candidate VUs for aggregation

On selecting a particular AU, a list of candidate river segments appears under “River segments to include in aggregation” and the common properties of those segments appear under “Common parameter values”. Two or more river segments must be selected to perform aggregation. The output folder in which the aggregated statistics will be written appears under “Output data”. Currently, the output folder is fixed to the same folder used for output of verification statistics. After defining one or more AUs, aggregation is performed by clicking “**Run**.”

Editing of the VUs upon which one or more AUs is based will result in a warning message and the option to either remove the edited VU from each of the AUs to which it belongs or to cancel the edits.

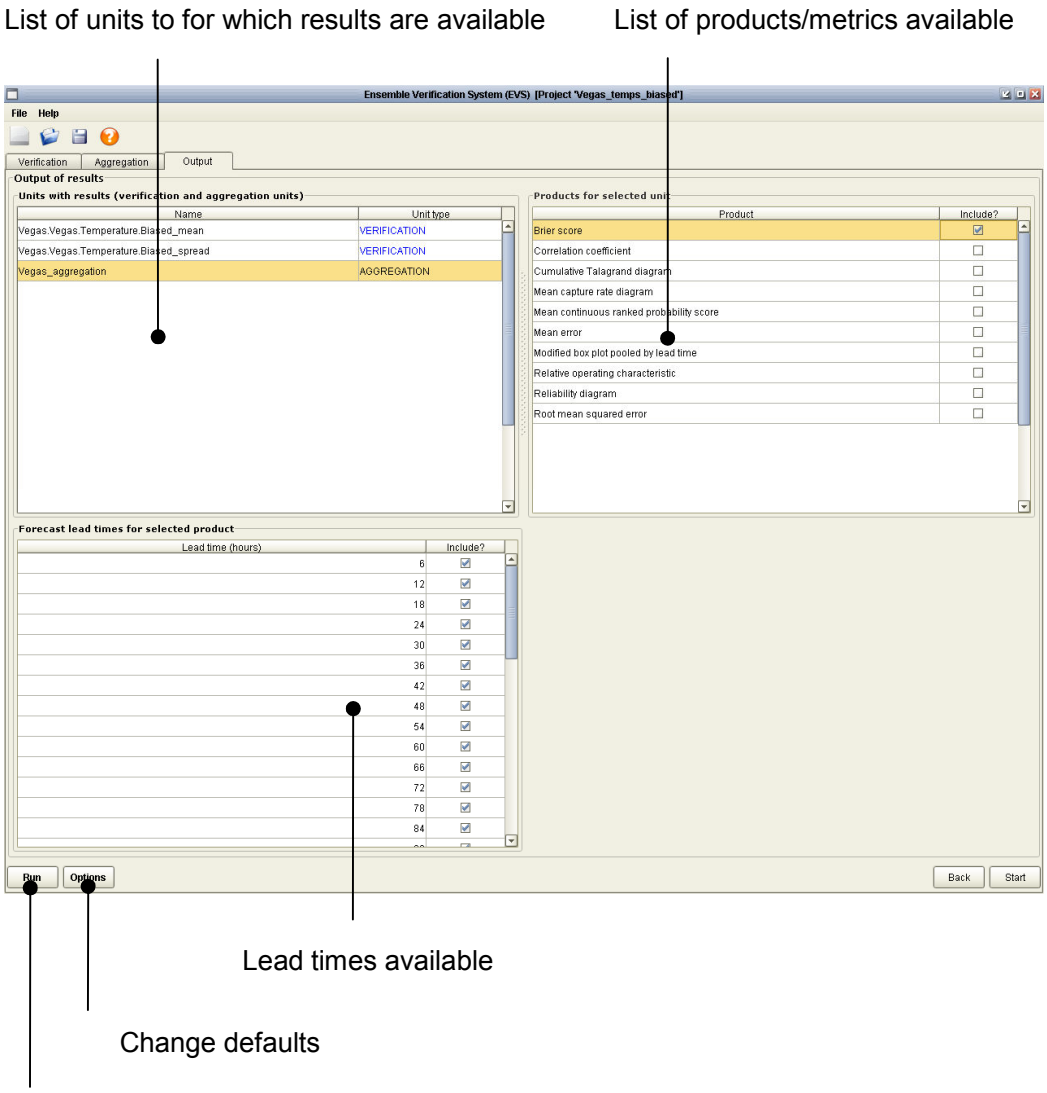
Display: window 1

The display section of the EVS allows for plotting of the verification statistics from one or more individual or aggregated verification units (i.e. one or more VUs and AUs). The units available for plotting are shown in the top left table, with VUs colored blue and AUs colored red (*figure 7*). On selecting a particular unit under “Units to plot”, a list of metrics with available results appears in the right-hand table. On selecting a particular unit, the bottom left table displays a list of lead times (in hours) for which the metric results are available.

When verifying or aggregating the paired data, the sample from which statistics are computed is generated by pooling pairs from equivalent lead times. Products may be generated for some or all of these lead times, and will vary with the metric selected. For example, in selecting 10 lead times for the modified box plot, it is possible to produce one graphic with 10 boxes showing the (pooled) errors across those 10 lead times. In contrast, for the reliability diagram, one graphic is produced for each lead time, with reliability curves for all thresholds specified in each graphic. These defaults may be extended in future. The units, products, and lead times and are selected by checking the adjacent boxes in the last column of each table. In addition, when the product and lead time tables are populated, right clicking on these tables will provide additional options for rapid selection of multiple products and lead times.

Products are generated with default options by clicking “**Run**”. The default options are to write the numerical results in XML format and the corresponding graphics in png format to the given output folder. The file naming convention is ‘`unit_identifiers_lead_time.file_extension`’ for plots that comprise a single lead time and ’`unit_identifiers.file_extension`’ for the plots that comprise multiple lead times and for the numerical results.

Figure 7: The first window in the Output pane



The default options for generating products are defined for each unit, and may be edited by selecting the "**Options**" button (*figure 8a/b*). For example, the numerical results and graphics may be plotted directly instead of, or in addition to, writing them. The image parameters and file formats can also be modified. When plotting results for multiple graphics in the internal viewer, a warning is given when more than five graphics will be plotted (in case a mistake was made). A tabbed pane is used to collect plots together for metrics that have one plot for each lead time (*figure 9*). For rapid viewing, these plots may be animated by pressing the "**Animate**" button.

Figure 8a: product writing options

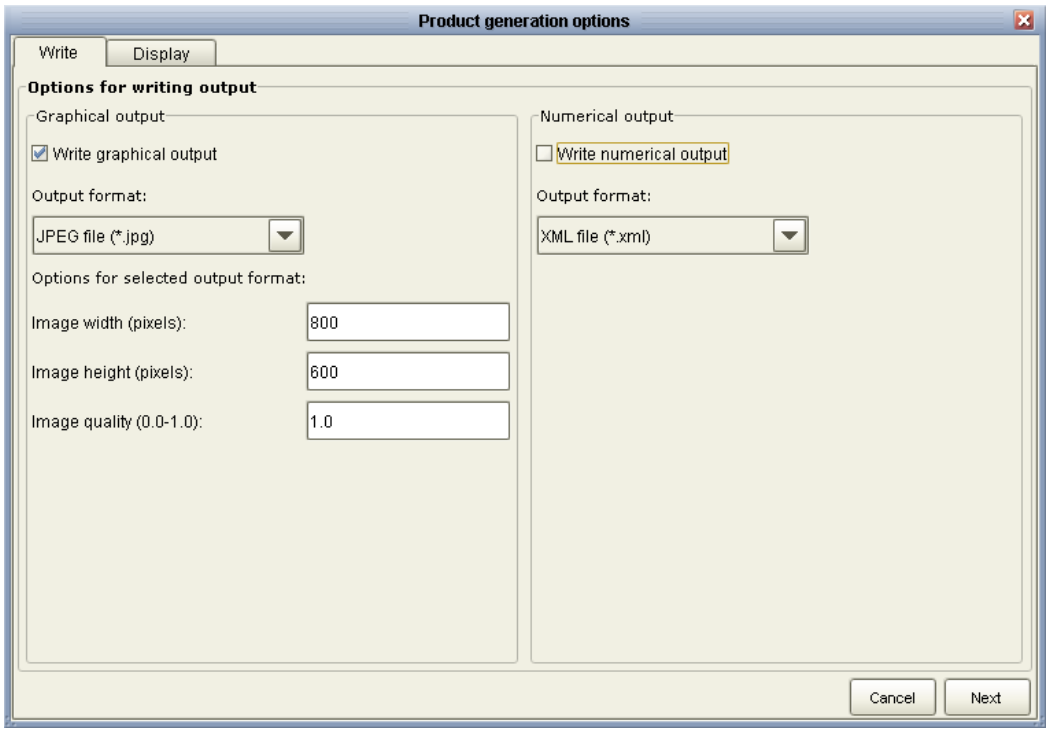


Figure 8b: product display options

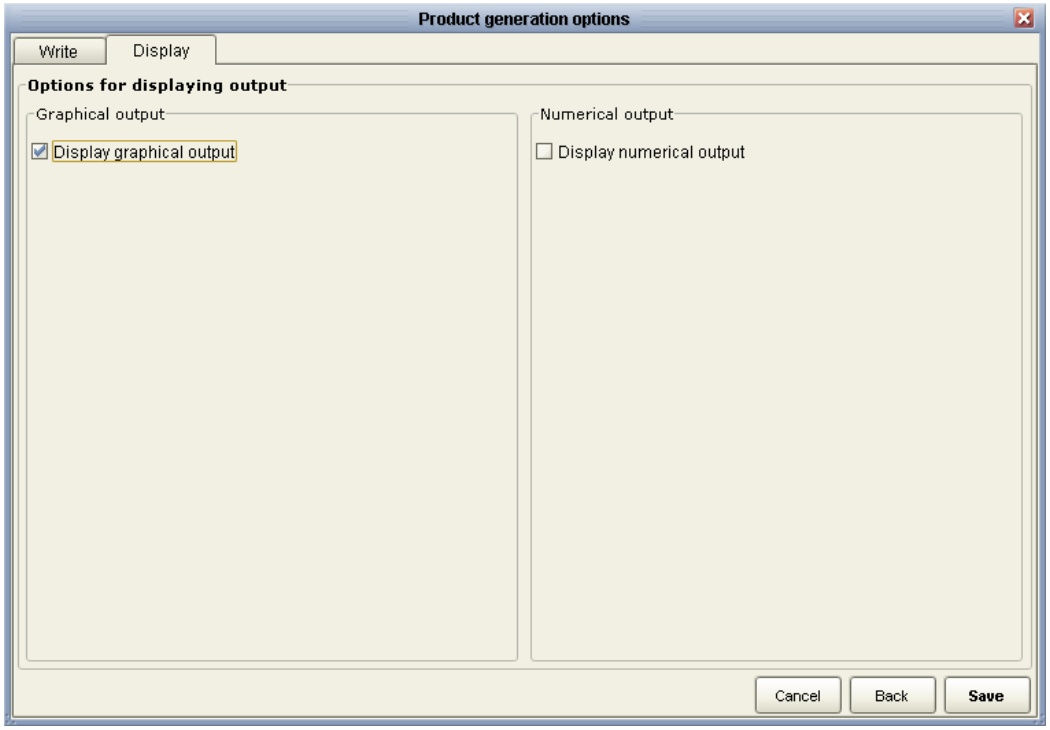
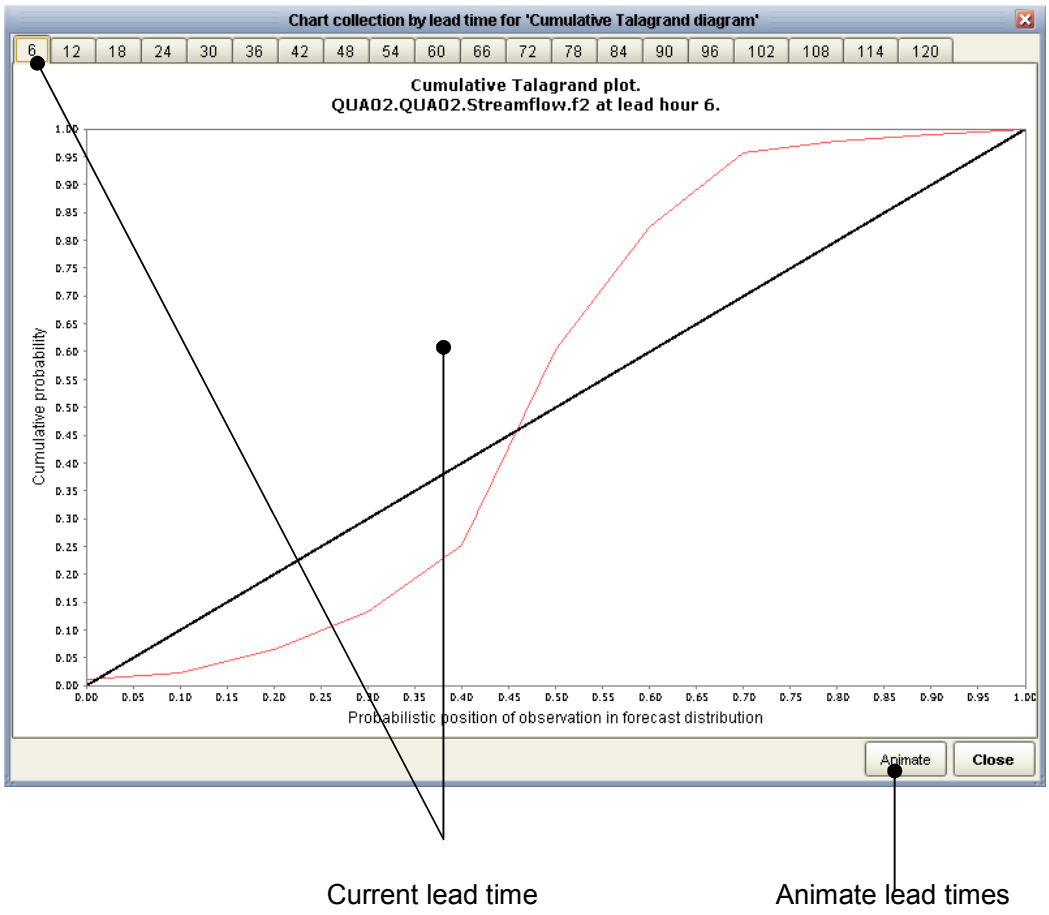


Figure 9: plot collection for a metric with one plot for each lead time

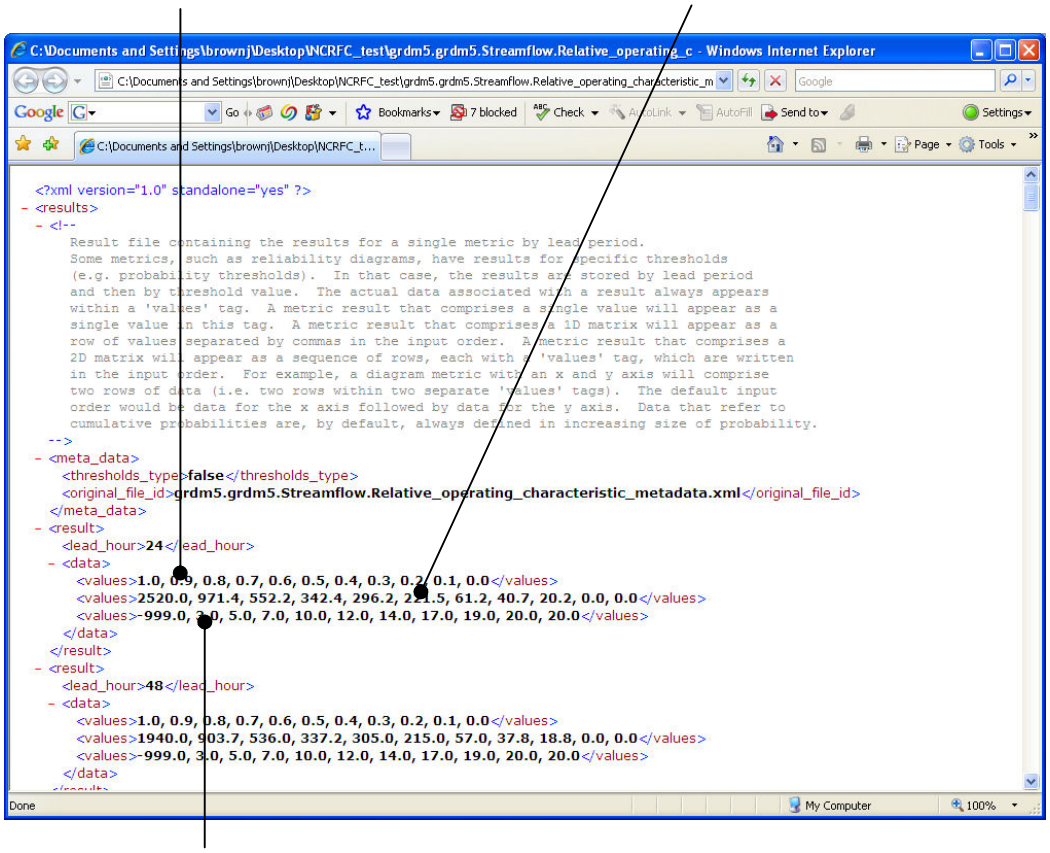


When writing numerical outputs for metrics that are based on one or more thresholds of the observations, such as the Brier Score, Relative Operating Characteristic and Reliability diagram, information about these thresholds is written to an XML file with the `_metadata.xml` extension. Specifically, the probability thresholds are written for each timestep, together with their values in real units (of the observations) and the numbers of samples selected by those thresholds. An example is given in *figure 10*.

Figure 10: example of a metadata file for metrics based on observed thresholds

Probability thresholds used at first lead time

Real values of thresholds



Sample counts for each threshold

APPENDIX A1 VERIFICATION STATISTICS COMPUTED IN EVS

EVS supports the following verification statistics:

- Deterministic error statistics for single-value forecasts, namely the ensemble means: Mean Error, RMSE, Correlation Coefficient;
- Brier Score (BS);
- Mean Capture Rate Diagram;
- Mean Continuous Ranked Probability Score (CRPS);
- Modified box plots;
- Reliability diagram;
- Relative Operating Characteristics (ROC);
- Cumulative Talagrand diagram.

Below is a short description of each metric, which is also available in the GUI (see *figure 4*).

Deterministic error statistics

Mean error

The mean error measures the average difference between a set of forecasts and corresponding observations. Here, it measures the average difference between the ensemble mean forecasts and observations.

The average error, $\overline{E}$, of n pairs of ensemble mean forecasts, $\overline{Y}$, and single-valued observations, x, is given by:

$$\overline{E} = \frac{1}{n}\sum_{i=1}^{n}(\overline{Y}_i - x_i)$$

The mean error provides a measure of first-order bias in the forecasts, and may be positive or negative.

Root Mean Squared Error (RMSE)

The Mean Squared Error (MSE) is the average squared forecasting error. The RMSE provides the square root of this value, which has the same units as the forecasts and observations (unlike the MSE). Here, the forecast is given by the ensemble mean value and an 'error' represents the difference between the forecast mean and the observation. For example, given two ensemble forecasts with mean

values 23.6 and 24.5 and corresponding observations, 22.1 and 22.2, the RMSE is given by:

$$\sqrt{\frac{(23.6-22.1)^2+(24.5-22.2)^2}{2}}=1.94$$

The general equation for the RMSE of n pairs of ensemble mean forecasts, $\overline{Y}$, and single-valued observations, x, is given by:

$$\mathrm{RMSE}(x,\overline{Y})=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\overline{Y}_i-x_i)^2}$$

The RMSE provides an indication of the 'average deviation' between the forecast value (in this case, the ensemble mean) and an observation in forecast units. The RMSE is always positive.

Correlation coefficient

The correlation coefficient measures the linear relationship between two variables. Here, it measures the linear relationship between n pairs of ensemble mean forecasts and corresponding observations. The maximum correlation coefficient is 1.0, which denotes a strong positive (linear) relationship between the forecasts and observations, or -1.0, which denotes a strong negative (linear) relationship (i.e. the observed values increase when the forecasts values decline and vice versa). It should be noted that the forecasts and observations may be perfectly correlated yet biased. In other words, a linear regression of the forecasts and observations would have a non-zero intercept on the y-axis. The minimum correlation coefficient is 0.0, which denotes no linear relationship between the forecasts and observations. It should also be noted that a low correlation coefficient may occur in the presence of a strong *non-linear* relationship, because the correlation coefficient measures linearity only.

EVS computes the Pearson's product-moment correlation coefficient, r, which is given by:

$$r=\frac{\mathrm{Cov}(x,\overline{Y})}{s_x\cdot s_{\overline{Y}}}$$

Where $\mathrm{Cov}(x,\overline{Y})$ is the sample covariance between the ensemble mean forecasts, $\overline{Y}_1....\overline{Y}_n$, and observations, $x_1 \ldots x_n$. The sample standard deviations of the

forecasts and observations are denoted $s_{\overline{Y}}$ and s_x, respectively. An unbiased estimate of the sample covariance between n pairs of forecasts and observations, $\mathrm{Cov}(x, \overline{Y})$, is given by:

$$\mathrm{Cov}(x, \overline{Y}) = \frac{1}{n-1} \sum_{i=1}^{n} \{(\mu_x - x_i)(\mu_{\overline{Y}} - \overline{Y}_i)\}$$

Where μ_x and $\mu_{\overline{Y}}$ are the sample means of the forecasts and observations, respectively. The sample mean of all the forecasts, $\mu_{\overline{Y}}$, should not be confused with the ensemble mean of an individual forecast, $\overline{Y}_i$.

Brier Score (BS)

The BS measures the average squared error of a probability forecast. It is equivalent to the mean squared error of a deterministic forecast, but the forecasts, and hence error units, are given in probabilities. The Brier Score measures the error with which a discrete event, such as 'flooding', is predicted. For continuous forecasts, such as the amount of water flowing through a river, one or more discrete events must be defined from the continuous forecasts. There are several ways in which an event may be defined, depending on the verification problem. However, a complete picture is only obtained by computing the BS for a representative range of events from the full distribution of forecasts and observations. For an event that involves exceeding some threshold, t, the Brier Score (or half Brier Score) is computed from n pairs of forecasts and observations:

$$\mathrm{BS} = \frac{1}{n} \sum_{i=1}^{n} \{(\mathrm{Prob}[Y_i > t] - \mathrm{Prob}[x_i > t])^2\}$$

Note that the observed probability is 0.0 if the event does not occur (x_i<=t) or 1.0 if the event does occur (x_i>t) at any given forecast time/location. A set of forecasts and observations match exactly in terms of BS if the mean squared difference between them is zero and hence BS=0.0.

Mean Capture Rate Diagram

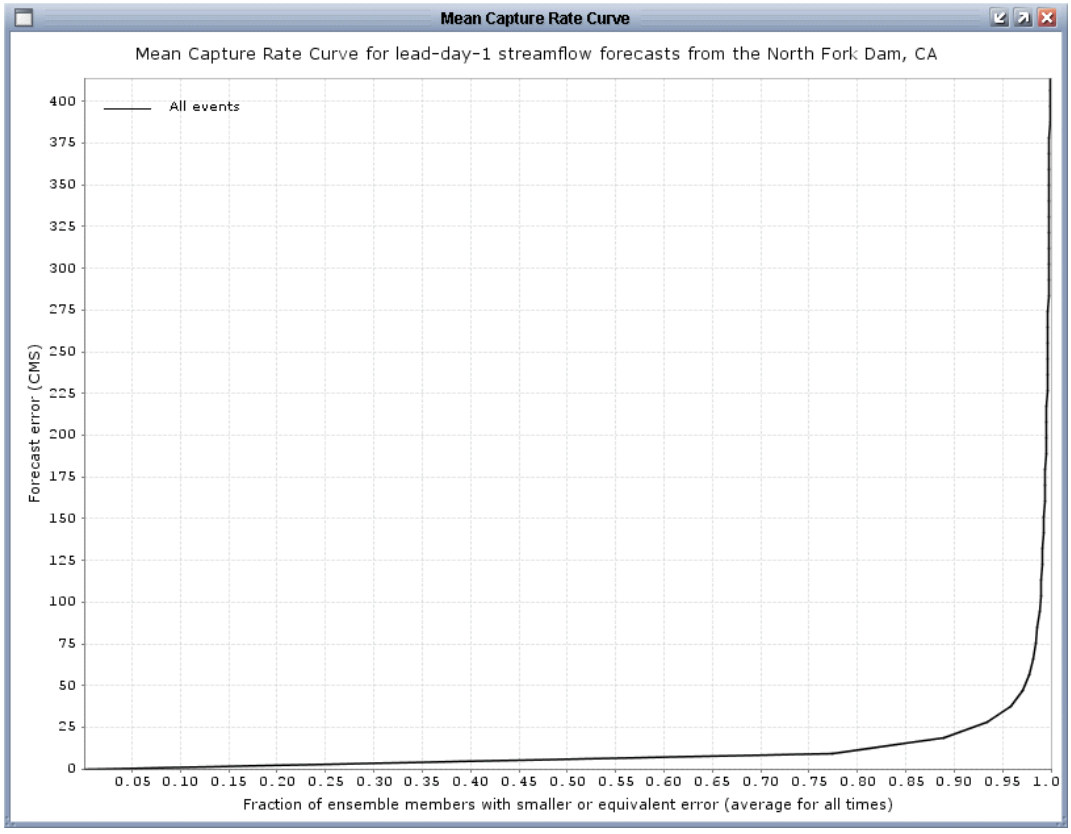
A key aspect of forecast quality is the probability of making a given error in real terms. The Probability Score (PS) of Wilson *et al*. (1999) is useful here because it identifies the probability with which a given 'acceptable error' is met. The concept of an 'acceptable error' is widely used in operational forecasting. The PS is defined for a symmetric window or 'acceptable error', w, around the observation, x:

$$PS(w) = F_Y(x + 0.5w) - F_Y(x - 0.5w) = \int_{x-0.5w}^{x+0.5w} f_Y(y)dy$$

It conveys the extent to which an observation is captured by the forecast, where a high capture rate implies greater forecast performance. The disadvantages of the PS include its subjectivity and sensitivity to hedging (Bröcker and Smith, 2007), whereby the expected value of the PS is maximized for sharp forecasts. By averaging the PS over a set of n ensemble forecasts and repeating for all possible windows, w, the probability of achieving a given acceptable error can be determined, hereafter referred to as the Mean Capture Rate (MCR):

$$\mathrm{MCR(w)} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{PS(w)} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}} \{\mathrm{F_Y(x_i + 0.5w) - F_Y(x_i - 0.5w)}\}$$

It should be noted that sensitivity to hedging does not apply to the MCR when evaluated for all w, as the result is not a score. The resulting curve may be separated into errors of over-prediction and under-prediction by computing the MCR for ensemble members that exceed the observation and fall below the observation, respectively. The MCR for 6-hourly forecasts of streamflow at the North Fork, CA, USA, are shown below:



The forecasts were evaluated at lead day 1 for the period 10/02/2000 to 09/29/2003. Deviations from the x-axis represent a declining capture rate; that is, an increasing probability of exceeding a given error or an increasing error for a given probability of capture. For example, there is a ~0.9 probability that a randomly selected ensemble member will have a forecast error less than or equal to 25 cms.

Mean Continuous Ranked Probability Score (CRPS)

The CRPS summarizes the quality of a continuous probability forecast with a single number (a score). It measures the integrated squared difference between the cumulative distribution function (cdf) of a forecast, $F_Y(y)$ and the corresponding cdf of the observations, $F_X(x)$.

The CRPS for a single forecast and observation is given by:

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} [F_Y(y) - F_X(y)]^2 \, dy$$

where $F_Y(y)$ is the cumulative probability distribution of the forecast and $F_X(x)$ is a step function that reaches probability 1.0 for values greater than or equal to the observation, and has probability 0.0 elsewhere. In practice, the CRPS is averaged across n pairs of forecasts and observations, which leads to the mean CRPS:

$$\overline{\mathrm{CRPS}} = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} [F_{Yi}(y) - F_{Xi}(y)]^2 \, dy$$

The numerical value of the CRPS will vary with application and is difficult to interpret in absolute terms (e.g. in terms of specific forecast errors). However, the CRPS has some desirable mathematical properties, including its insensitivity to hedging (i.e. the expected value of the score cannot be improved, *a priori*, by adopting a particular forecasting strategy). Other scores, such as the Probability Score of Wilson *et al*. (1999), may be hedged against (in this case by issuing sharper forecasts).

Modified box plots

Box plots (or box-and-whisker diagrams) provide a discrete representation of a continuous empirical probability distribution (Tukey, 1977).

Building on this idea, an empirical pdf may be summarized with an arbitrary set of percentile bins of which an arbitrary proportion may be shaded (e.g. the middle 60%), to convey the outer and inner probability densities, respectively. The percentile bins

are specified as probabilities in the probability threshold parameter box (e.g. 0.1 represents a percentile bin of 0-10%).

Reliability diagram

The reliability diagram measures the accuracy (or bias) of the forecast probabilities. According to the reliability diagram, the probability with which an 'event' is forecast should match the probability with which it is observed, for all possible events. This is a sufficient condition for unbiasedness of the forecast probabilities, and implies that the marginal distributions are also identical. For continuous numerical variables, one or more events may be defined using probabilities of exceedence. In other words, a probability of 0.9 will produce a reliability diagram for the top 10^{th} percentile of forecast events. The Reliability diagram plots the forecast probabilities on the x-axis against the (conditional) observed probabilities for a single forecast event on the y axis.

In order to compute the reliability diagram, a dichotomous event must be defined, such as Y > t, where t is a flood threshold. The forecasts are then grouped into n bins that exhaust the unit interval according to Prob[Y>t], each containing m forecast-observation pairs. The average probability within each bin is used as the plotting position on the x axis. Thus, for one bin we have:

$$\frac{1}{m}\sum_{i=1}^{m}\text{Prob}[Y_i > t]$$

The number of forecasts within each bin, m, is referred to the sharpness of the forecasts and is typically displayed as a histogram for all n bins alongside the reliability diagram, since a forecast may be very reliable without being sharp (e.g. climatology). Within each bin, the fraction of observations that meet the condition is then computed. Thus, for one bin we have:

$$\frac{1}{m}\sum_{i=1}^{m}I(x_i > t)$$

where $I(\cdot)$ is the indicator function (i.e. has value 1 if the condition is met, 0 otherwise). If the forecast is perfectly reliable, the observed fraction within each bin will equal the average of the forecast probabilities and the reliability diagram will contain a diagonal line. Deviation from the diagonal line represents bias in the forecast probabilities of a given event when the event is predicted with a given probability. The reliability diagram may be computed for several events.

Relative Operating Characteristic

The Relative Operating Characteristic (also known as the Receiver Operating Characteristic) measures the quality of a forecast for a dichotomous event that is predicted to occur (e.g. rainfall or flooding). It does not consider the quality of forecasts that predict no event (e.g. no rainfall or no flooding). The ROC diagram plots:

- X-axis: the probability that an observation, x, does not exceed a real-valued threshold, t, *when it is forecast to exceed that threshold with a given probability*, p_t, (probability of false detection or false positive rate, POFD(t,p_t)), repeated for several probability thresholds. In this case, each probability threshold will produce m forecast-observed pairs:

$$\mathrm{POFD}(t, p_t) = \frac{1}{m}\sum_{i=1}^{m}\{\mathrm{Prob}[x_i <= t] \,|\, (\mathrm{Prob}[Y_i > t] > p_t)\}$$

  Note that $\mathrm{Prob}[x_i <= t] \,|\, (\mathrm{Prob}[Y_i > t] > p_t)$ will assume the value 1 or 0.

- Y-axis: the probability that an observation, x, does exceed a real-valued threshold, t, *when it is forecast to exceed that threshold with a given probability*, p_t, (probability of detection or true positive rate, POD(t,p_t)), repeated for the same probability thresholds as used above:

$$\mathrm{POD}(t, p_t) = \frac{1}{m}\sum_{i=1}^{m}\{\mathrm{Prob}[x_i > t] \,|\, (\mathrm{Prob}[Y_i > t] > p_t)\}$$

  Note that $\mathrm{Prob}[x_i > t] \,|\, (\mathrm{Prob}[Y_i > t] > p_t)$ will assume the value 1 or 0.

These values are computed for probability thresholds that exhaust the unit interval, which is normally defined by a number of plotting points, n.

For a forecast to perform well in terms of ROC, the probability of detection must be high relative to the probability of false detection. A forecasting system that produces random forecasts in line with climatological expectation will have as many successful predictions of an event as unsuccessful ones. Hence, a skillful forecasting system will always produce a ROC curve that lies above the diagonal line.

Cumulative Talagrand diagram

A simple method for assessing the reliability of a set of ensemble forecasts is to count the number of observations that fall within a particular region of the forecast distribution. To be perfectly reliable, $F_Y(y)$ should always capture x, and x should fall within any given probabilistic window, p_w, of $F_Y(y)$ in proportion to the size of p_w. For example, by defining p_w in relation to the forecast median and computing the reliability, $REL(p_w)$, over m ensemble forecasts and corresponding observations we have:

$$REL(p_w) = \frac{1}{m}\sum_{i=1}^{m} I(0.5 - 0.5p_w \le F_Y(x)_i \le 0.5 + 0.5p_w)$$

where $I(\cdot)$ is the indicator function (i.e. has value 1 if the condition is met, 0 otherwise). The forecast is perfectly reliable, with respect to window p_w, if $REL(p_w) = p_w$. Anchoring the window to the center of the forecast distribution is sensible if the tails are subject to large sampling uncertainties. By computing $REL(p_w)$ for an exhaustive set of probability windows (e.g. deciles), the overall reliability of the forecasts can be determined. This is analogous to the Talagrand diagram (rank histogram, multi-category reliability diagram; Anderson, 1996; Hamill, 1997; Talagrand, 1997), only defined with probability windows, p_w, rather than ranked ensemble members. It should be noted that $REL(p_w)$ provides a weaker definition of reliability than the conventional reliability diagram (Hsu and Murphy, 1986; Wilks, 1995), which tests the conditional probabilities, $F_{X|Y}(x|y)$, for all possible events, y. However, it is easier to construct and interpret, and experience points to a good correlation between the two.

Research-level metrics (not accessible via GUI)

Before a metric becomes available for operational use in EVS (and, therefore, accessible via the GUI), it may be trialed. These research metrics are accessible only via additional commands in the EVS project file. They cannot be accessed via the GUI, although the results for these metrics will be appended to the list of metrics in the Output dialog.

Currently, there are two additional metrics available for research purposes. Both are variants of the Modified box plot (see above). By default, the Modified box plots in EVS are computed for forecasts that are pooled by forecast lead time. Two non-pooled versions of the modified box plots are available for research purposes. First, it is possible to construct a modified box plot whereby each box represents a single forecast (at a given lead time) ordered from the first forecast valid date. There will be one box plot for each lead time. Secondly, it is possible to construct a modified box

plot whereby each box represents a single forecast (at a given lead time) ordered by the size of the observed value. These box plots are useful for identifying individual blown forecasts. Examples are given below.

Modified box plots were derived from 6-hourly forecasts of temperature and precipitation at Huntingdon, PA, USA. The temperature forecasts were evaluated at lead day 1 for the period 12/1993-12/1998. The precipitation forecasts were evaluated at lead day 1 for the period 08/2002-09/2005. The plots were generated for deciles of the empirical pdfs, of which the middle 60 percent were shaded. *Figure A1_1* shows the temperature errors against forecast date, $F_E(e)$ or $F_{E|T}(e|t)$, while *figure A1_2* shows precipitation errors against observed values, ordered by increasing magnitude of precipitation and for positive observed precipitation only, $F_{E|X}(e|x)$ where $x > 0$. The boxes are averaged for duplicate values of x. These plots reveal a range of conditional biases in the temperature and precipitation forecasts. For example, the latter consistently under predict large observed events (*Figure A1_2*). This is consistent with the calibration of meteorological models for 'average conditions'.

Figure A1_1: Temperature forecast errors against forecast time at Huntingdon, PA

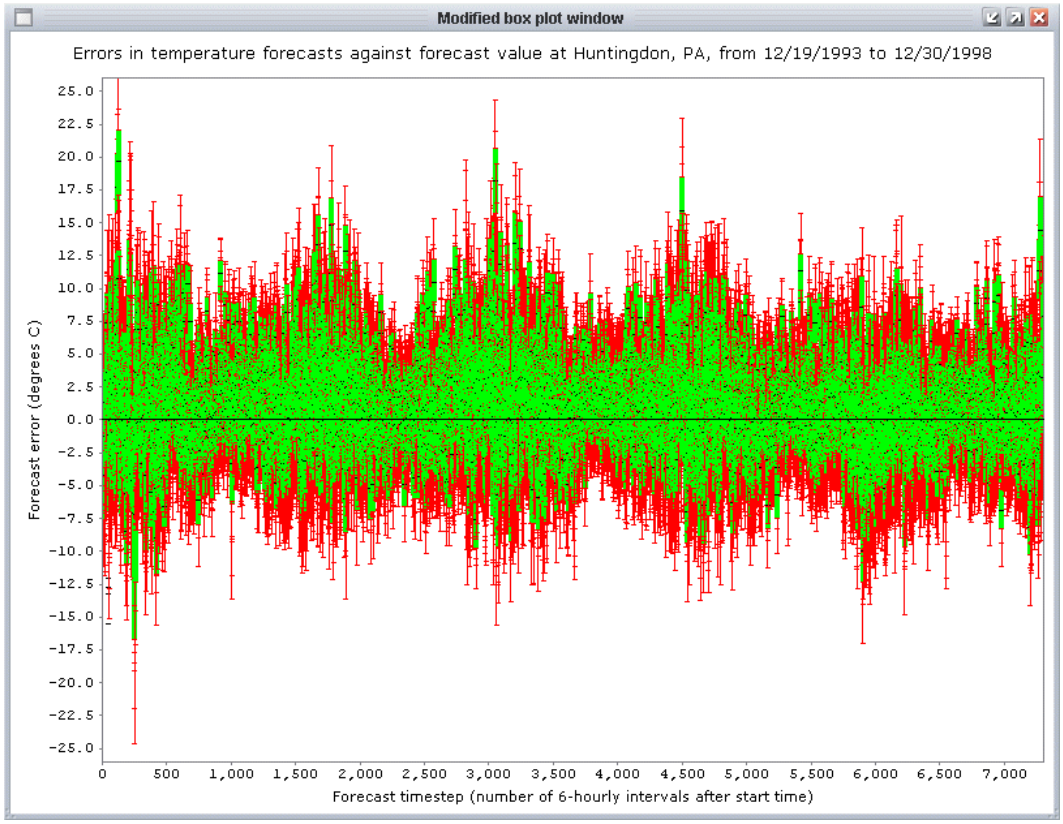
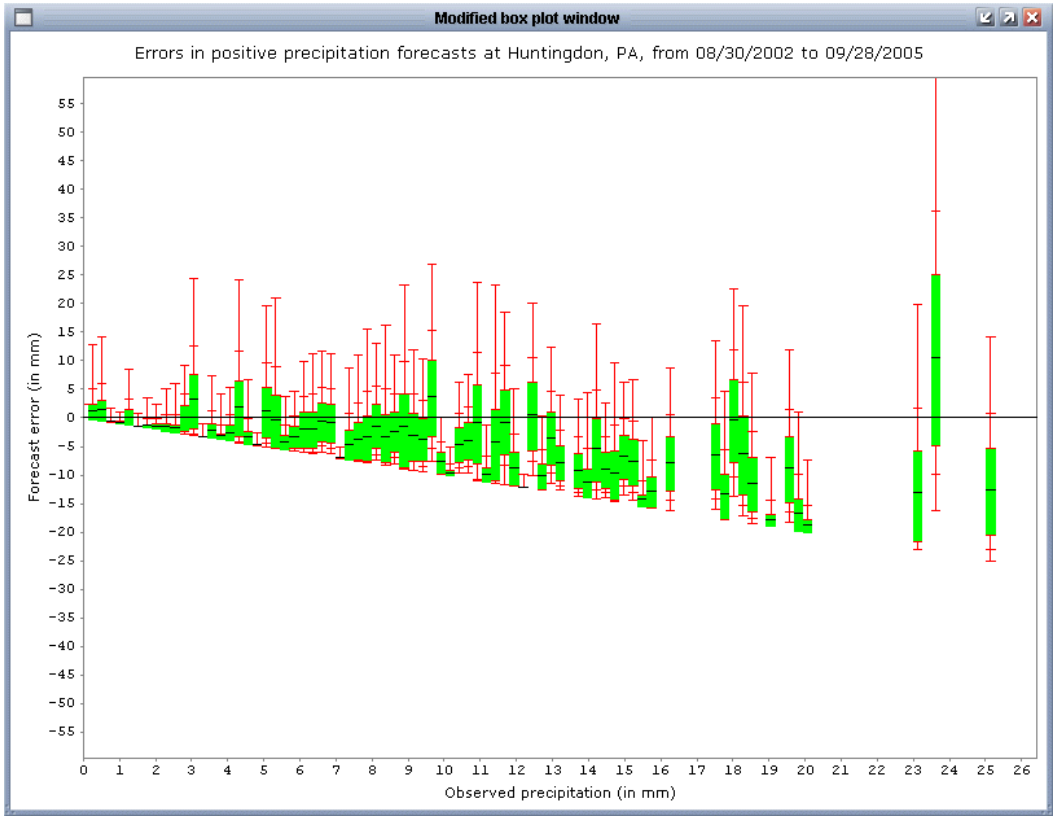


Figure A1_2: Errors in forecast precipitation vs. observed value at Huntingdon, PA



These plots are obtained by adding the following information to an existing EVS project file under the metrics section of that file. This information must be added under the metrics section of *every* Verification Unit for which the metrics are required. The parameters have the same meaning as those for the default Modified box plot (see *table 2*). See *Appendix A2* for information about the structure of the project file.

<u>The information to add for plots ordered by time and size of observation, respectively:</u>

```
<metric>
       <name>ModifiedBoxPlotUnpooledByLead</name>
       <box_unpooled_lead_points_parameter>10</box_unpooled_lead_points_param
       eter>
       <forecast_type_parameter>regular</forecast_type_parameter>
       <unconditional_parameter>false</unconditional_parameter>
</metric>


<metric>
       <name>ModifiedBoxPlotUnpooledByLeadObs</name>
       <box_unpooled_lead_obs__points_parameter>10</box_unpooled_lead_obs_poi
       nts_parameter>
       <forecast_type_parameter>regular</forecast_type_parameter>
       <unconditional_parameter>false</unconditional_parameter>
</metric>
```

APPENDIX A2 INPUT AND OUTPUT DATA

Input Data

For each verification unit:

1. Observation file: 6-hr observed values in the datacard file with 1 value per line for a specific time series ID. Any file name may be used, although EVS searches for .OBS file extensions by default.

   Examples of observed data files:

   - precipitation: <time_series_id>.MAP06.OBS
   - temperature: <time_series_id>.MAT.OBS
   - streamflow: <segment_id>.<time_series_id>.QINE.06.OBS

2. Ensemble forecast files for a range of dates (referred as mm/dd/yyyy) for a specific time series ID. Again, the file naming convention is unimportant.

   Examples of forecast data files:

   - precipitation: datacard file with 4 values per line called <yyyymmdd><time_series_id>.MAP06
   - temperature: datacard file with 4 values per line called <yyyymmdd><time_series_id>.MAT
   - streamflow: the CS file generated by ESP

3. Climatology file [optional]. Same format as the observed data file.

Output Data

EVS produces three types of file, namely: 1) project files, which store previously defined VUs and AUs; 2) paired data files, which store the paired forecasts and observations associated with a single VU; and 3) product files, such as the numerical and graphics results associated with a particular verification metric.

Project files

Project files store all of the parameters required to close and restart EVS without loss of information. A project file is produced or updated by clicking "**Save**" or "**Save as…**" at any point during the operation of EVS. The data are stored in XML format

and are, therefore, human readable, and may be produced separately from EVS (e.g. for batch calculations in the future).

The XML contains the following tags, in hierarchical order:

Level 1 (top level):

<verification> //Top level tag
<verification_unit> //Tag for a single verification unit (see *Level 2*)
<aggregation_unit> //Tag for a single aggregation unit (see *Level 3*)

Level 2 (verification unit, VU):

<verification_unit>
    <identifiers> //Identifiers for the VU (see *Level 2a*)
    <input_data> //Input data, including forecasts and observations (see *Level 2b*)
    <verification_window> //Verification window (see *Level 2c*)
    <output_data_location> //Path to output data folder
    <paired_data_location> //Path to paired data file [only when defined]
    <metrics> //Verification metrics selected (see *Level 2d*)

Level 2a (VU identifiers):

<identifiers> //Identifiers for the VU
    <river_segment> //Identifier for the verification point (river segment)
    <time_series> //Identifier for the time-series
    <environmental_variable> //Variable id (e.g. streamflow)
    <additional_id> // Additional id (e.g. forecast_model_1) [only when defined]

Level 2b (VU input data sources):

<input_data> //Identifiers for the VU
    <forecast_data_location> //Forecast data
        <file> //Path to first file/folder (e.g. first file in a file array or a folder)
        <file> //Path to second file in a file array [only when defined]
        <file> //Etc.
        …
    <observed_data_location> //Path to observed data file
    <forecast_time_system> //Name of forecast time system
    <observed_time_system> //Observed time system
    <climatology_time_system> //Climatology time system [only when defined]
    <forecast_support> //Scale of forecasts

<statistic> //E.g. “instantaneous”
<period> //E.g. “1” [only when defined: blank when statistic = instantaneous]
<period_units> //E.g. “DAY” [only when defined: as above]
<attribute_units> //E.g. “cubic feet/second”
<notes> //Additional textual info. [only when defined]
<forecast_support> //Scale of observations [see forecast_support]
<climatology_support> //Scale of climatological obs. [see forecast_support]

Level 2c (verification window for a given VU):

<verification_window> //Window parameters
<start_date> //Start date (in forecast time system)
<year> //Start year
<month> //Start month of year
<day> //Start day of month
<end_date> //See start date
<forecast_lead_period> //Maximum forecast lead period considered
<forecast_lead_units> //Units for the maximum lead period
<aggregation_lead_period> //Average X consec. leads U [only when defined]
<aggregation_lead_units> //Period units for averaging (U) [only when defined]
<date_conditions> //Date conditions (see *Level 2c_1*) [only when defined]
<value_conditions> //Value conditions (see *Level 2c_2*) [only when defined]

Level 2c_1 (date conditions on the verification window) [only when defined]:

<date_conditions> //Date conditions
<exclude_years> //Integer years to exclude from the overall range
<exclude_months> //Integer months to exclude from the overall range
<exclude_weeks> //Integer weeks to exclude from the overall range
<exclude_days_of_week> //Integer days to exclude from the overall range

Level 2c_2 (value conditions on the verification window) [only when defined]:

<value_conditions> //Value conditions.
<condition> //First of n possible conditions
<unit_id> //Identifier of the VU on which the condition is built
<forecast_type> //True for forecasts, false for observed values
<statistic> //Name of statistic, e.g. mean
<consecutive_period> //Moving window size [only when defined]
<consecutive_period_units> //Moving window time units [only when defined]

<logical_conditions> //Set of n possible logical arguments
<function> //First logical argument
<name> //Unary function name, e.g. isLessThan (<)
<value> //Unary function threshold, e.g. 0.5 means "< 0.5"
…
…

Level 2d (verification metrics for a given VU):

<metrics> //Set of n possible metrics to compute
<metric> //First of n metrics
<name> //Name of metric
Storage of parameters follows: varies by metric
…

Level 3 (aggregation unit, AU) [only when defined]:

<aggregation_unit> //Aggregation unit
<name> //The aggregation unit name
<unit_id> //First of n possible VU identifiers associated with the aggregation unit
…
<output_data_location> //Path to where output data should be written for the AU

Paired data files

A paired data file stores the pairs of forecasts and observations for a single VU in XML format. The file name corresponds to the VU identifier with a `_pairs.xml` extension.

Each pair comprises one or more forecasts and one observation, and is stored under a `<pr>` tag. Each pair has a readable date in Coordinated Universal Time (UTC or GMT), a lead time in hours (`<ld_h>`), an observation (`<ob>`), one or more forecast values (`<fc>`), and an internal time in hours (`<in_h>`) used by EVS to read the pairs (in preference to the UTC date). The internal time is incremented in hours from the forecast start time (represented in internal hours) to the end of the forecast lead period. When multiple forecasts are present, each forecast represents an ensemble member, and each ensemble member is listed in trace-order, from the first trace to the last. An example of the first few lines of a paired file is given below:

<pr> //First pair
<dt> //Date tag
<y>2005</y> //Year

```
            <m>11</m>   //Month
            <d>31</d>   //Day
            <h>18</h>   //Hour
      </dt>   //End of date tag
      <ld_h>6.0</ld_h>   //Lead time in hours
      <ob>150.625</ob>   //Observed value
      <fc>   //Forecast values: in this case 49 ensemble members
            157.31567,157.31598,157.31627,157.3342,157.3148,
            157.31598,157.31509,157.31509,157.31572,157.31567,
            157.31538,157.31598,157.31598,157.3148,157.31627,
            157.31393,157.31567,157.31598,157.31595,
            157.31627,157.32852,157.31569,157.3148,157.34517,
            157.34586,157.34148,157.31664,157.31538,
            157.31509,157.31644,157.31509,157.31567,
            157.31639,157.31598,157.31598,157.31627,
            157.31598,157.31567,157.3161,157.31538,157.34439,
            157.3148,157.31627,157.3148,157.31598,157.31598,
            157.31657,157.3156,157.31567
      </fc>
      <in_h>315570</in_h>   //Internal hour incremented from start time
</pr>    //End of first pair tag
```

.....

Product files

Product files include the numerical and graphical results associated with verification metrics.

Currently, the graphical files are written in one of two formats, namely the Joint Photographic Experts Group format (`.jpeg` extension) and the Portable Network Graphic format (`.png` extension).

Numerical results are written in XML format. One file is written for each metric. The file name comprises the unique identifier of the VU or AU, together with the metric name (e.g. `Aggregation_unit_1.Modified_box_plot.xml`). Some metrics, such as reliability diagrams, have results for specific thresholds (e.g. probability thresholds). In that case, the results are stored by lead period and then by threshold value. The actual data associated with a result always appears within a 'values' tag. A metric result that comprises a single value will appear as a single value in this tag. A metric result that comprises a 1D matrix will appear as a row of values separated by commas in the input order. A metric result that comprises a 2D matrix will appear as a sequence of rows, each with a <values> tag, which are written in the input order. For example, a diagram metric with an x and y axis will comprise two rows of data (i.e. two rows within two separate <values> tags). The default input order would be data for the x axis followed by data for the y axis. Data that refer to cumulative probabilities are, by default, always defined in increasing size of probability. If

available, sample counts are given in the last <values> tag. Sample counts are also printed out in a separate XML file for each threshold used in the ROC, Reliability and Brier Score metrics (thresholds are compulsory for these metrics). This information is written to a file with the VU identifier, metric name and a `_sample_counts.xml` extension.

An example of the first few lines of a numerical result file for one metric, namely the 'modified box plot', is given below:

```
<meta_data>     //Tag for metadata on the results

            //Next tag indicates that results are not available for separate
            thresholds of the observed distribution

            <thresholds_type>false</thresholds_type>
            <original_file_id>Aggregation_unit_1.Modified_box_pl
            ot.xml</original_file_id > //Original file
</meta_data>   //End of metadata
<result>   //First of n possible results
      <lead_hour>6</lead_hour>     //Result applies to lead hour 6
      <data>   //Start of data
            <values>0.0,0.1,…</values>   //Probs. drawn in box diagram
            <values>-1102,-233.5,…</values>   //Real values of probs.

            ……
      </data>   //End of data
</result>   //End of first result
……
```

APPENDIX A3 REFERENCES

Anderson, J. L. (1996) A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, **9**, 1518 1530.

Elmore, K.L. (2005) Alternatives to the Chi-Square Test for Evaluating Rank Histograms from Ensemble Forecasts. *Weather and Forecasting*, **20(5)**, 789-795.

Hamill, T.M. (1997) Reliability diagrams for multicategory probabilistic forecasts. *Weather and Forecasting*, **12**, 736-741.
Hamill, T.M. (2001) Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, **129**, 550-560.

Hersbach, H., (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15**, 559-570.

Hsu, W.-R. and Murphy, A.H. (1986) The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, **2**, 285-293.

Joliffe, I.T. and D. B. Stephenson, (ed), 2003: Forecast Verification, A Practitioners Guide in Atmospheric Sciences, Wiley, West Sussex, England.

Murphy, A.H. (1973) A new vector partition of the probability score. Journal of Applied Meteorology, **12**, 595-600.

Talagrand O. (1997) Assimilation of observations, an introduction. *Journal of the Meteorological Society of Japan*,**75**, 191–209.

Tukey, J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, MA. 688pp.

Wilson, L.J., Burrows, W.R. and Lanzinger, A. (1999) A strategy for verification of weather element forecasts from an ensemble prediction system. *Monthly Weather Review*, **127**, 956-970.

World Meteorological Organization (WMO), 2004: WWRP/WGNE Joint Working Group on Verification, Forecast Verification – Issues, Methods and FAQ, web site: http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html