

**THIS LOG CONTAINS A LIST OF CHANGES BETWEEN SUCCESSIVE VERSIONS OF THE EVS**

---

FROM EVS 2.0 (build date 10/01/09) to EVS 3.0 (build date 10/01/10).

\*\*\*\*\*

\* Feature upgrades and modifications not related to Graphical User's Interface \*

\*\*\*\*\*

- Packaged the EVS into a single executable JAR file with all associated libraries using an ANT build script. The EVS is now delivered as a single executable JAR file, without the need to maintain an internal directory structure for dependent libraries (which are now packaged and accessed from within the executable JAR).
- The writing of conditional pairs (i.e. a subset of the overall pairs with any conditions on variable value or date applied) has been made optional to speed-up the processing of large numbers of verification points. This option is implemented via the `<write_conditional_pairs>` tag of the `<paired_data>` section of the EVS project file, with a default of `true`, i.e. conditional pairs are written by default, as before. This option is also accessible via the GUI (see below).
- Improved the performance of the temporal aggregation routine and provided options for the type of aggregation function applied not only to the forecast data, but to the forecast valid times and lead times. Also changed the default behavior; previously, the forecast valid time was given the mean of the input times and the lead time assumed the maximum of the input times; the forecast valid time now assumes the maximum of the input times (see below).
- Improved information and error messages printed to standard out (i.e. the console, if EVS was initiated from a console window).
- Added options for aggregating the support of the observed data to match the support of the forecasts, including the ability to compute an accumulation over a forecast window. Eventually, an aggregation function will be implemented for every possible combination of input support allowed in the EVS (including a change of measurement units).
- Added methods for reading paired data in the same ASCII format to which the XML pairs can be converted. Thus, paired files may be produced and read in ASCII format (as well as forecasts and observations) where convenient.
- Implemented additional R scripts for plotting the verification results produced by the EVS (i.e. the XML output), including a script that will plot the EVS single-valued metrics and ensemble scores as a function of threshold value.
- Implemented the three-part decomposition of the Brier Score into:  $\text{Brier Score} = \text{reliability} - \text{resolution} + \text{uncertainty}$  and the associated graphical and numerical products.
- Implemented additional options for averaging the forecast ensemble members when computing single-valued verification metrics such as the mean error, root mean square error, mean absolute

error, and correlation coefficient. The default remains to compute the ensemble mean. Additional options now include the ensemble median and mode.

- Added the Mean Absolute Error of the ensemble average to the single-valued verification metrics.
- Added the climatological frequency and the zero-skill line to the reliability diagram (located half-way between the climatological frequency and the expected frequency for a reliable forecasting system). These curves are available in the XML output files, but are not plotted within the EVS.
- Backwards compatibility has been maintained for earlier project files. Upon saving old project files within a new version of EVS, new options will be written with their default values.
- Implemented the concept of “main” and “auxiliary” thresholds for metrics that either require or support thresholds. Currently, the “main” thresholds are used to identify events (or subsets of data) that should be included in the graphical outputs from the EVS. By default, both the “main” and “auxiliary” thresholds are included in the numerical outputs from the EVS. This information is stored as an additional XML tag in the project file, labeled `main_threshold`, which contains a list of Boolean values equal in length to the number of thresholds (`true` indicates a main threshold).
- Added the facility to derive climatological probability thresholds from a larger set of observed data. In the EVS Version 2.0 they were derived from the paired observations, after applying any requested changes in units, temporal aggregation, or value and date conditions, including the discrete verification time-period requested in the first verification window. Now, they may be derived from the original observed data, again after applying any requested changes in units, temporal aggregation, or date and (observed) value conditions, but NOT the discrete verification time-period (i.e. the full period of record covered by the observed file will be used, after applying any changes in measurement units, temporal aggregation, value conditions on the observations, and date conditions *except* for the reduced verification time period). This option is controlled by a Boolean flag, which is accessible via a check box in the GUI, and also in the project file under the new XML tag, labeled `use_all_observations_for_climatology`. Note that this option only applies to the derivation of real-valued thresholds corresponding to particular climatological probabilities of occurrence. For those metrics that incorporate climatological probabilities in the calculation (e.g. the climatological frequency in the reliability diagram), the behavior is unchanged (the observations associated with the conditional pairs are still used).
- Added further date and value (pre-)conditions, including additional statistics for selecting forecasts based on value (ensemble median and mode, probability of not exceeding a given value, and the value corresponding to a given non-exceedence probability) and additional functions for selecting forecasts and observations based on dates (hours of day in UTC). Thus, much more complex pre-conditions are now possible, such as selecting only those pairs (for verification) whose forecast probability of exceeding “flood stage” is greater than 0.9. This functionality is necessary for real-time verification, where the aim is to select historical (observed and forecast) analogs to a real-time forecast based on specific properties of that real-time forecast, and possibly auxiliary information.
- Added the facility to compute the binormal approximation to the Relative Operating Characteristic (ROC) curve and the associated ROC Score. The approximation is based on fitting the binormal model to the empirical (POD, POFD) pairs and is, therefore, dependent on the number of thresholds

chosen. For an exact comparison between the binormal approximation to the ROC curve and the binormal approximation to the ROC Score, a common number of thresholds should be adopted for each metric. However, when comparing the empirical ROC Score to the binormal ROC Score, the results will be closest when adopting  $m+1$  thresholds, where  $m$  is the number of ensemble members per forecast. Specifically, the empirical ROC Score is derived from ranking of the POD and POFD data, rather than computing the ROC curve. The ranked data can take at most  $m$ , “jumps” in probability (at the corresponding ensemble member positions). Hence, the empirical ROC Score is analogous to deriving the ROC Score from an empirical ROC curve constructed with  $m+1$  thresholds.

- Added the facility to aggregate the observed support prior to verification. Previously, this was only possible for the forecasts. The same restrictions apply to change of support of the observations as the forecasts, namely the aggregated support is exactly divisible by the frequency of the data and comprises either a mean of the input values if the inputs have instantaneous support or a total of the input values if the input values are totals.
- Added the option to remove certain lead times from the verification results based on a minimum sample-size requirement. The sample size constraint is set by a fraction in the range  $[0,1]$ . The fraction is multiplied by the average number of pairs across all lead times to determine the minimum sample size in numbers of pairs. For example, a fraction of 0.5 implies that verification results will *not* be computed for any lead time with fewer than 50% of the average number of pairs across all lead times.

The following new features are only accessible via the EVS project file (not the GUI):

- Added the facility to specify the method for computing CRPS in the EVS project file. By default the Hersbach (2000) method is used, but a method that can handle null ensemble members has been added. This is specified in the `<crps_method>` tag of the crps metric in the EVS project file, with options `hersbach` and `with_nulls`. If null members are present, the `hersbach` option will lead to all forecasts with one or more null members being removed from the calculation, and will fail to compute if all forecasts contain one or more null members.
- The facility to prevent elimination of duplicate pairs (pairs with a common forecast valid time and lead time) has been added to the EVS project file. This is implemented via the `<eliminate_duplicates>` tag of the `<paired_data>` section of the EVS project file, with a default value of `true`. When `false`, duplicate pairs will not be eliminated. This is necessary when computing verification metrics for data that have been pre-pooled across several forecast locations and contained in a *single* paired file.
- Added an option to pool the verification pairs across several forecast locations and to compute the verification metrics from the pooled pairs, rather than averaging the metrics from the individual locations. Theoretically, this approach is preferred, but is much more time-consuming in practice, and is usually not feasible. The default behavior remains to average the verification metrics from the individual locations. The new option is only accessible via the EVS project file by adding or

setting the `<pool_pairs>true</pool_pairs>` entry to the XML for a particular aggregation unit, where `true` implies that pooling will be conducted, and `false` implies averaging.

- Added the facility to change the decimal writing precision of the paired data for a given verification unit. This applies to both the raw and conditional pairs. The default behavior is to write forecasts and observations with a maximum precision of five decimal places (unchanged), with fewer places written as required. The integer number of decimal places ( $>0$ ) can now be defined in the project file using the `<paired_write_precision>5</paired_write_precision>` tag, which is part of the `<paired_data>` block. This functionality is not accessible via the GUI, and existing pairs will not be re-written with a new decimal precision (unless re-writing is otherwise necessary).
- Added the facility to set the behavior for removing null ensemble members when writing the paired file. The default behavior remains to *remove* null ensemble members. This may be changed using the `<strip_nulls_from_paired_file>true</strip_nulls_from_paired_file>` tag, which is part of the `<paired_data>` block. This functionality is not accessible via the GUI, and existing pairs will not be re-written (unless otherwise required).

```
*****  
*                               Changes in default behavior                               *  
*****
```

- Changed the temporal aggregation default to store the maximum valid time of the input times (in UTC) when conducting aggregation. Previously, the default was to compute the mean of the input times. The default for handling the forecast lead times remains to compute the maximum of the inputs. Thus, for example, aggregation of four six-hourly pairs at increasing UTC times of {18, 0, 6, 12} previously generated an aggregated paired value with time UTC 3, but will now generate an aggregated paired value with time UTC 12 (note that 12 proceeds 18 when considering date). Thus, the aggregated value should be interpreted as the value over the period of aggregation immediately preceding the stated time.
- Changed the start and end dates of the verification period defined in the first verification window from the forecast time zone to UTC. The start and end dates begin and end at 00 UTC on the specified date, respectively. Thus, in order to include verification pairs that fall on the specified end day, one day should be added to the input date.
- Changed the order of error messages displayed in the GUI Console window (not to be confused with an external console) so that the latest error messages are displayed at the top of the console rather than appended to the bottom.

```
*****  
*                               Feature upgrades and modifications for developers                               *  
*****
```

- Added numerous additional methods for developers that assist in sub-setting paired data according to varied conditions met in particular rows or columns of the paired-data matrix. These methods can be found in `evs.utilities.matrix.DoubleMatrix2D`. The conditions can be made



- Added advanced options for computing different averages from the ensemble members when using single-valued verification metrics.
- Added auto-recall of the last directory accessed when creating, saving and reading project files so that the last working directory is opened by default.
- Added an option to change the behavior for writing conditional pairs. This is controlled by a check button in the advanced options (accessed via the “More” button) in the “Output” section of the first window in the GUI. The default (slower) behavior is to write conditional pairs.
- Added support for multiple-row selection and deletion in the thresholds table associated with each verification metric (in the second verification window).
- Added the facility to distinguish between “main” and “auxiliary” thresholds for metrics that either require or support thresholds. By default, the “main” thresholds are used for plotting and all thresholds (“main” and “auxiliary”) are added to any numerical outputs written by the EVS.
- Added functionality to generate verification thresholds semi-automatically for verification metrics that either require or support thresholds. The thresholds are generated by entering a number of thresholds, the first threshold value, and a constant increment (positive or negative) between thresholds. This is useful for designing plots that show verification scores as a ‘continuous’ function of threshold value (i.e. outside of the EVS). By default, only those thresholds identified as “main” thresholds are included in the graphical outputs from the EVS, but the numerical outputs (on which custom plots are designed) include all of the thresholds.
- Re-labeled the “Edit no-data value” option in the advanced input data options dialog to the more generic “Edit other options” and added a check box to control the way observed data are used to determine climatological probability thresholds.
- Improved the GUI for selecting pre-conditions to apply to the verification pairs, in keeping with the enhanced functionality for identifying pre-conditions (see above).
- Added option to iconify (or “minimize” in Windows terminology) the EVS GUI while processing a verification project; a button labeled “iconify” has been added to the progress dialog to facilitate this.
- Changed the label `Forecast lead period` to `Forecast lead time horizon` in the first window of the GUI.
- Added the binormal approximation to the ROC and ROC Score to the GUI. In both cases, the binormal approximation is appended to the results when selecting to do so under the Advanced Parameter Options dialog. In that case, the empirical ROC data are plotted as open points and the binormal approximation is plotted with a line of the same color.
- Removed the text (`in forecast time system`) for both the start and end dates of the verification period in the first window of the GUI, reflecting the change to UTC (see above).
- Added an error message when attempting to temporally aggregate forecasts over a longer period than the specified forecast lead time horizon (e.g. attempting to compute monthly averages for forecasts with a lead time horizon of 14 days).
- Improved the error console in the GUI.

\*\*\*\*\*

\* Bug fixes related to Graphical User's Interface \*

\*\*\*\*\*

- Fixed a bug in the table of thresholds for each verification metric, ensuring that the scrolling window expands properly as new thresholds are added (previously a fixed limit).
- Fixed a bug in the table of candidate units for aggregation, ensuring that the scrolling window expands properly to show all available units (previously a fixed limit).
- Corrected a bug in the GUI for selecting pre-conditions to apply to the verification pairs. Entering incorrect conditions lead to a (correct) warning, but, when subsequently cancelling further edits, the existing (valid) conditions were removed rather than being returned to the original (valid) state.



FROM EVS 1.0 (build date 05/09/08) to EVS 2.0 (build date 10/12/09).

\*\*\*\*\*

\* Feature upgrades and modifications not related to Graphical User's Interface \*

\*\*\*\*\*

- Added multiplication factor in support dialog to allow simple changes between measurement units (more complex operations, such as a change in temperature units, are not yet supported).
- Implemented reading of PI-XML observations
- Implemented reading of PI-XML forecasts
- Implemented reading of ASCII observations
- Implemented reading of ASCII forecasts
- Changed representation of forecast lead times from integer hours to double-precision float hours to allow verification of forecasts with lead times in fractional hours, thereby extending the EVS to arbitrary forecast lead times.
- Rewrote the online documentation and updated the mathematical formulas for all of the verification metrics.
- Implemented an R script for each metric in the EVS to read in the XML output and produce high quality plots in EPS format for scientific papers.
- Modified calculation of the mean CRPS to account for the relative position of the observation between the two adjacent ensemble members.
- Added ROC score to the available metrics and included a plot by forecast lead time (same as with other scores). The calculation is based on Mason and Graham (2000).
- Added a sample size metric and associated plot to compute the basic sample size information by forecast lead time and threshold. This may be used for exploratory purposes before computing other verification metrics. In future, we may add further metrics for data exploration (of the observed and forecast data rather than the verification pairs).
- Added modified box plot by size of observed value to GUI (previously via the command line only).
- Modified the spatial aggregation routine to compute the expected (mean) value of each metric across a set of Verification Units rather than pooling paired data.
- Included ability to perform a weighted spatial aggregation. The weight is uniform by default and must sum to 1. A non-uniform weight is also permissible and a weight of "S" is used to weight by the sample size at the first lead time (i.e. maintaining constant weights across lead times). If a verification metric is not available for a given lead time or threshold the weights are automatically rescaled to sum to 1, maintaining the correct relative weighting of the available metrics.
- Relaxed constraints on spatial aggregation to allow aggregation for Verification Units with different start and end dates.
- Improved the efficiency of file reading for external file formats to ensure that only data within the specified start and end dates and forecast lead times are fully read (otherwise only the file headers are read to check this information).

- Implemented backwards compatibility for the above features so that they do not prevent running of old EVS projects. However, the aggregation routine and CRPS update has been swapped without the option of backwards compatibility. Thus, old projects with spatial aggregation will produce different results in the EVS 2.0. The CRPS update was a bug fix, voiding the need for backwards compatibility.
- Updated the algorithm for CRPS to the method described in Hersbach, H., 2000: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. There are small numerical differences between the old and new algorithms. Also, the procedure described in Hersbach assumes that a constant number of ensemble members is available, whereas the previous method for computing CRPS had no such constraint. Thus, differences will be seen when comparing numbers between systems for which some forecasts comprise null ensemble members.
- Added the decomposition of the CRPS into reliability, resolution and uncertainty, as described in Hersbach (2000).
- Added the Brier Skill Score (BSS) for an arbitrary reference forecast selected by the user.
- Added the Continuous Ranked Probability Skill Score (CRPSS) for an arbitrary reference forecast selected by the user.

\*\*\*\*\*

\* Feature upgrades and modifications related to Graphical User's Interface \*

\*\*\*\*\*

- Removed Time-Series ID and renamed River Segment ID to Location ID.
- Moved basic output options from pop-up window to main Output window.
- Implemented enhanced error dialog with improved error messages.
- Implemented enhanced progress monitor to monitor and display progress of paired-file reading (and included a pair count in the paired file to enable this).
- Updated the aggregation window to include a weighting input in the table of verification metrics. Also renamed some features in this window.
- Added the option to select an arbitrary reference forecast for a skill metric.
- Added the option to show skill score decompositions in a tabbed pane (similar to metrics with one plot per lead time), which may be animated.
- Improved display of zero error line in plots (extended continuously).
- Improved auto-scaling of axes in plots.



FROM EVS 1.0 BETA (build date 10/12/07) to EVS 1.0 (build date 05/09/08).

\*\*\*\*\*

\* Feature upgrades and modifications not related to Graphical User's Interface) \*

\*\*\*\*\*

- Allowed real-valued thresholds for all metrics.
- Allowed probability thresholds for all metrics, not just Brier, ROC, and Reliability.
- Included the option for thresholding with a closed interval (i.e. a "between" condition).
- Supported the use of symmetric windows around the forecast median in the Spread-Bias plot.
- Included sharpness (sample-count) plot in the Reliability diagram.
- Changed the definition of probability thresholds in Spread-Bias plot, Mean Capture Rate diagram and Box Plots. Previously, these thresholds referred to plotting positions (i.e. plot resolution) and NOT thresholds of the observed distribution. They now refer to thresholds of the observed distribution for consistency with all other metrics. Plotting positions are now determined with a 'points count' parameter. For example, a point count of 10 for the Spread-Bias plot will construct a plot comprising 10 points.
- Added a new 'points count' parameter for the Spread-Bias plot, Mean Capture Rate diagram, Box Plots and Reliability diagram, which allows the resolution of those diagrams to be altered.
- Included the option to change the default temporal aggregation function from the mean over a specified period to one of several other functions, including the minimum, maximum, and total (i.e. accumulation).
- Included units in the plots that comprise real units (mean error, RMSE, Mean Capture Rate diagram, box plots) once those units have been added to the observed and forecast support for a verification unit.
- Included an option to animate a sequence of verification graphics at different lead times.
- Included writing of sample counts to an XML file when writing other numerical results.
- Included writing of conditional pairs to XML format as well as the original pairs.
- Included an option to ignore global value conditions on a per-metric basis. For example, if a condition was applied to consider only those pairs whose ensemble mean temperatures exceeded freezing, this condition could be ignored on a per-metric basis.
- Enabled backwards compatibility with old project files (i.e. projects with old options will run as before).
- Enhanced and updated documentation.
- Improved memory management for Aggregation units.

\*\*\*\*\*

\* Feature upgrades and modifications related to Graphical User's Interface) \*

\*\*\*\*\*

- Removed the table containing reference forecasts, which are not yet enabled.

- Improved the labeling of various options (e.g. 'time zones' rather than 'time systems').
-