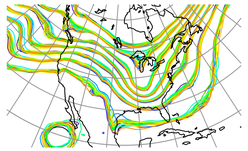

A Quantile Conserving Ensemble Filtering Framework:
Next Generation Nonlinear and Non-Gaussian Data Assimilation Capabilities for DART

Jeffrey Anderson

NCAR;CISL;Data Assimilation Research Section, Boulder, CO

dart@ucar.edu

1. The Data Assimilation Research Testbed (DART)

The Data Assimilation Research Testbed (DART) is a community facility for ensemble data assimilation developed and maintained at the National Center for Atmospheric Research (NCAR). DART provides data assimilation (DA) capabilities for nearly all NCAR community earth system models and many other models.

The DART website (https://dart.ucar.edu) includes documentation, tutorials, and examples of DART use and a list of DART publications and presentations.

2. Original DART Ensemble DA Algorithms

The schematic in Fig. 2.1 shows how DART implements ensemble DA. The default DART algorithms assume a normal distribution for step III (this is the ensemble adjustment Kalman filter; EAKF) and use linear regression for step IV. This poster discusses quantile conserving ensemble filters for step III that can assume arbitrary distributions for the prior and the observation error, and probit transformed quantile regression methods for step IV that allow much more general regression for computing state increments.

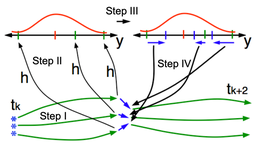

Figure 2.1: The DART assimilation framework. A model makes an ensemble of forecasts (green arrows) valid at the next time at which observations will be assimilated in step I. Observations for that time are assimilated one at a time. The expected values of an observation for each ensemble (green ticks, top left) are computed using a forward operator function (h) in step II. Ensemble increments for the observed quantity (blue vectors, upper right) are computed in step III. Step IV linearly regresses the observation increments onto each state variable; increments for all state variables can be computed in parallel.

3. Quantile Conserving Ensemble Filters

Quantile conserving filters provide a very general method of computing increments for the prior ensemble of an observed quantity (step III, Fig. 2.1). The algorithm is illustrated in Figs. 3.1 and 3.2. This example uses a simple normal distribution which reproduces the existing DART EAKF. Figure 3.3 shows the impact of using different continuous prior PDFs for the same ensemble. Figure 3.4 shows an example for a doubly bounded quantity (sea ice fractional coverage is a physical example) using a beta prior PDF and a beta observation likelihood. Quantile conserving filters are especially useful for bounded quantities like tracer concentrations, depths of things like snow or ice, and estimating model parameters that have a restricted range. Figures 3.3 and 3.4 are reproduced from Anderson (2022, https://doi.org/10.1175/MWR-D-21-0229.1) which also provides a table of dozens of continuous PDF families for which it is straightforward to compute the posterior PDF.

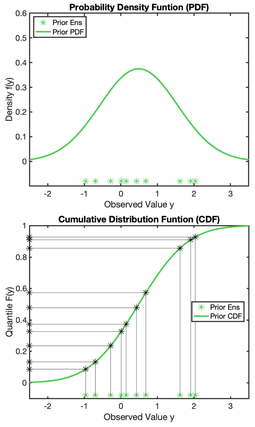

Figure 3.1: Part I of a quantile conserving ensemble filter for a continuous normal probability distribution function (PDF). In the top panel, a continuous PDF (green curve) is fit to a prior observation ensemble (green asterisks). In the bottom panel, the corresponding cumulative distribution function (CDF, green curve) is used to compute the quantile (black asterisks) of each ensemble member.

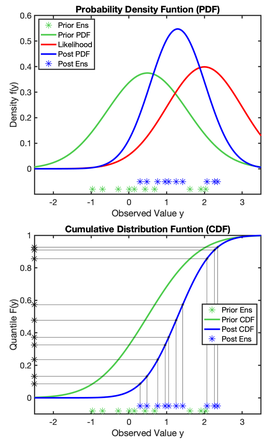

Figure 3.2: Part II of a quantile conserving ensemble filter. In the top panel, the product of the continuous prior PDF and a continuous observation likelihood (red curve) is computed to give the continuous posterior PDF (blue curve). In the bottom panel, the corresponding posterior CDF (blue curve) is used to invert the prior quantiles to get the posterior ensemble (blue asterisks).

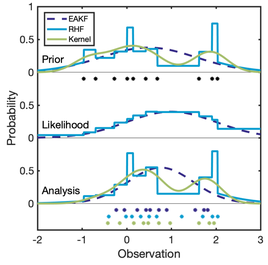

Figure 3.3: A 10-member prior ensemble is indicated by the black asterisks below the top panel. Continuous prior (top) and analysis (bottom) distributions for this ensemble are shown for an EAKF, a rank histogram filter (RHF), and a kernel filter with kernel variance set to 10% of the ensemble variance. (middle) A continuous normal likelihood used for the EAKF and the kernel filter with mean 1 and variance 1, and the RHF piecewise constant approximation are shown. Asterisks mark the analysis ensembles for the EAKF, RHF, and kernel filter below the analysis (from top to bottom, respectively). In this example the ensemble is roughly normal so all methods give similar analysis ensembles.

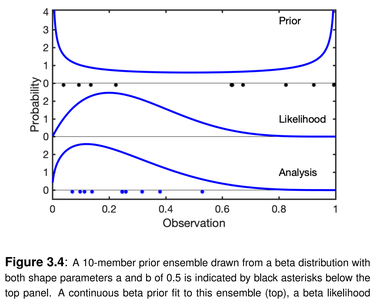

Figure 3.4: A 10-member prior ensemble drawn from a beta distribution with both shape parameters a and b of 0.5 is indicated by black asterisks below the top panel. A continuous beta prior fit to this ensemble (top), a beta likelihood with $\alpha = 2$ and $b = 5$ (middle), and the corresponding analysis beta PDF (bottom). Asterisks mark the analysis ensemble below the analysis distribution. The fifth and sixth smallest ensemble members have nearly the same values and are difficult to distinguish.

4. Probit Transformed Quantile Regression

While quantile conserving algorithms for step III lead to significant improvements in analysis estimates for observed variables, those improvements can be lost when using standard linear regression of observation increments to update other state variables in step IV. However, doing the regression of observation quantile increments in a probit-transformed, bivariate, quantile space guarantees that the posterior ensembles for state variables also have all the advantages of the observation space quantile conserving posteriors. For example, if state variables are bounded, then posterior ensembles will respect the bounds. The posterior ensembles also respect other aspects of the continuous prior distributions.

The new algorithm computes the quantiles of each ensemble member for both the observed variable and state variable (same as for observed ensemble in Section 3). The quantiles are bounded between 0 and 1. The quantiles are then further transformed to an unbounded domain by applying the probit function, the inverse of the standard normal CDF. The regression of the observation increments is done in this transformed space. The transformations are then inverted to get the posterior ensemble of the state variable.

Figure 4.1 illustrates some of the problems of using linear regression when a state variable has a distribution that is not normal. In this case, the state variable is non-negative, but linear regression leads to negative posterior ensemble members. Figure 4.2 shows how the new method improves the posterior ensemble. Figure 4.3 compares linear regression to the new method for a binormal state variable distribution.

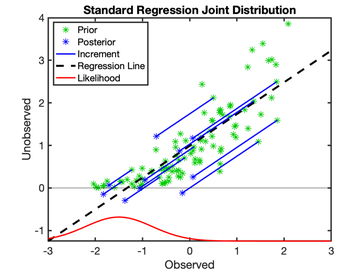

Figure 4.1: Green asterisks show an 80-member ensemble drawn from a bivariate distribution with a normally distributed observation (horizontal axis) and a gamma distributed state variable (non-negative, vertical axis). The red normal PDF is an observation likelihood. The least squares line for the ensemble is dashed black and increments using linear regression are shown in blue for 10 selected ensemble members. The regression can lead to meaningless negative tracer concentration.

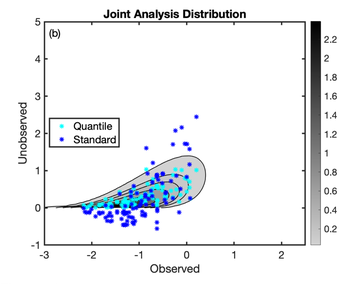

Figure 4.2: The posterior ensemble for the prior in Fig. 4.1 for standard regression (blue asterisks) and probit transformed quantile regression (cyan asterisks) superimposed on the shaded posterior PDF. The probit transformed quantile regression respects the bounds and is more consistent with the entire posterior.

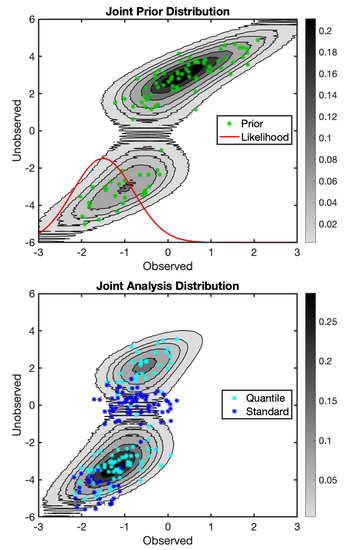

Figure 4.3: Prior (top) and posterior (bottom) ensembles for a normally distributed observed variable and a binormally distributed state variable. Standard regression results in many posterior members that are in neither high probability area; the new method is much more consistent with the posterior.

Inflation and localization, methods that improve the quality of ensemble DA, can also negate the advantages of the quantile conserving method. However, both localization and inflation can be done in the probit transformed quantile space as shown in Figs. 4.4 and 4.5.

Combining these new methods for steps III and IV can significantly improve data assimilation for non-Gaussian quantities in Earth system models.

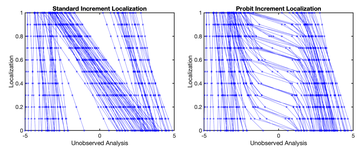

Figure 4.4: The result of applying increment localization for the state variable in Fig. 4.3. For both panels, the 80-member prior ensemble is plotted for localization 0, and the unlocalized posterior ensemble for localization 1. Standard DART localization (left) can lead to many ensemble members near 0; between the two high probability areas; see for instance the ensemble for localization 0.5. Localizing the increments in the probit transformed quantile space (right) avoids this deficiency. Ensemble members 'skip across' the low probability region as localization is increased.

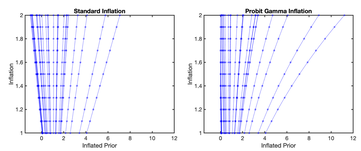

Figure 4.5: The result of applying inflation to the prior state ensemble in Fig. 4.1. The standard DART inflation results in negative ensemble members that make no sense (left). Applying inflation in the probit transformed quantile space avoids negative members while still generating additional spread (right).

5. DA in an Idealized Tracer Transport Model

DART now includes a low-order model of tracer transport for testing DA algorithms for bounded quantities. The model is an extension of the Lorenz-96 model that has been widely used for DA evaluation. Each gridpoint has the standard Lorenz-96 state, plus a tracer concentration and a tracer source/sink. A multiple of the state values are treated as a wind field at each gridpoint and used to advect the tracer. Figure 5.1 shows a time series of the wind field and Fig. 5.2 shows the corresponding tracer concentration. There is a single time constant tracer source at gridpoint 1 and a smaller sink at all other gridpoints. The wind is more often positive than negative and plumes of tracer propagate to the right across the domain. Sometimes the wind reverses leading to shorter plumes heading to the left. The tracer concentration is often exactly zero in some parts of the domain far from the source.

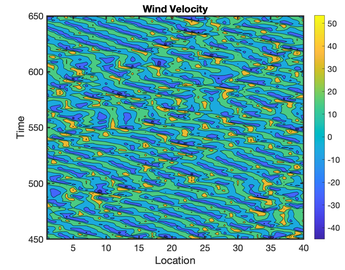

Figure 5.1: Time series of idealized winds from the Lorenz-96 tracer advection model. These winds are just the value of the standard state variables multiplied by a constant.

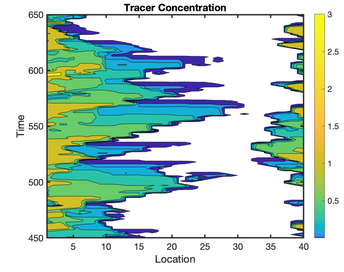

Figure 5.2: Time series of tracer concentration from the Lorenz-96 tracer advection model. There is a constant point source of tracer at gridpoint 1 and a constant small sink at all other points. Tracer is advected using an upstream semi-Lagrangian scheme with the winds in Fig. 5.1.

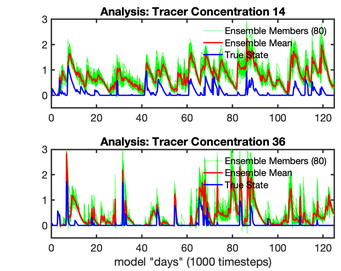

Figure 5.3: A time series of the ensemble analysis estimate of tracer at two gridpoints using DART's standard EAKF. The state (wind) and tracer concentration are observed at each grid point.

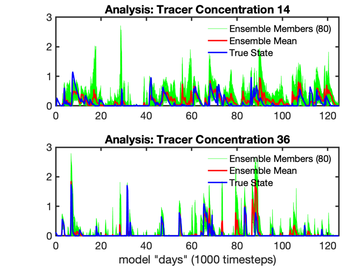

Figure 5.4: As in Fig. 5.3, but using a quantile conserving filter and probit transformed quantile regression. All prior distributions are bounded normal rank histograms. This method avoids bias due to the bounded nature of the tracer and is able to have all ensemble members go to zero when appropriate.

6. Try It

These new algorithms extend the capabilities of ensemble DA to general non-Gaussian and nonlinear distributions. Transformative improvements in DA can result for many applications. Largest improvements are found for bounded variables like tracer concentration, snow and ice depth, soil moisture, and similar quantities in other parts of the Earth system. Model parameters can also be estimated with DA and large improvements can occur for bounded parameters. Variables that have distinctly non-Gaussian prior distributions can also see large improvements. Examples can include atmospheric quantities like moisture and cloud amount in the presence of convection, and many land surface variables.

A DART release, that includes the algorithms described here and the idealized tracer model, is now available at the QR code (https://github.com/NCAR/DART/releases/tag/v11.0.0-alpha). Contact **dart@ucar.edu** with questions or if you would like to explore collaborations using these new algorithms. A description of the quantile conserving ensemble filter is available at the MWR link in Section 3 and in https://docs.dart.ucar.edu/en/quantile_methods/README.html.

7. Acknowledgements

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of NSF.

We would like to acknowledge high-performance computing support from Cheyenne (doi:10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.