A Nonlinear Rank Regression Method for Ensemble Kalman Filter Data Assimilation

JEFFREY L. ANDERSON

Data Assimilation Research Section, National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 27 December 2018, in final form 28 March 2019)

ABSTRACT

It is possible to describe many variants of ensemble Kalman filters without loss of generality as the impact of a single observation on a single state variable. For most ensemble algorithms commonly applied to Earth system models, the computation of increments for the observation variable ensemble can be treated as a separate step from computing increments for the state variable ensemble. The state variable increments are normally computed from the observation increments by linear regression using the prior bivariate ensemble of the state and observation variable. Here, a new method that replaces the standard regression with a regression using the bivariate rank statistics is described. This rank regression is expected to be most effective when the relation between a state variable and an observation is nonlinear. The performance of standard versus rank regression is compared for both linear and nonlinear forward operators (also known as observation operators) using a low-order model. Rank regression in combination with a rank histogram filter in observation space produces better analyses than standard regression for cases with nonlinear forward operators and relatively large analysis error. Standard regression, in combination with either a rank histogram filter or an ensemble Kalman filter in observation space, produces the best results in other situations.

1. Introduction

Anderson (2003) shows that it is possible to describe many varieties of ensemble Kalman filters without loss of generality by simply describing the impact of a single observation on a single state variable. In particular, the classical perturbed observation ensemble Kalman filter (EnKF; Burgers et al. 1998; Houtekamer and Mitchell 1998), which is perhaps more accurately referred to as a stochastic EnKF (Vetra-Carvalho et al. 2018), the ensemble adjustment Kalman filter (EAKF; Anderson 2001), and other similar square root filters (Tippett et al. 2003) can be described in this way. The description can be simplified further by separating the computation of increments for the prior ensemble estimate of an observation from the computation of increments for a state variable given increments for the observation. For the standard EnKF and EAKF, the second step simply linearly regresses observation increments onto state variable increments using the prior bivariate ensemble of the observation and the state variable.

Subsequent studies have explored other methods for computing observation increments while continuing to use linear regression to compute state increments. For example, the rank histogram filter (RHF; Anderson 2010) is a nearly nonparametric method for computing observation increments. The prior ensemble is used to generate an approximate continuous prior distribution for the observation by partitioning the real line into bins containing equal probability mass, conceptually similar to the rank histogram (Anderson 1996; Hamill 2001).

The gamma/inverse gamma and inverse gamma/gamma filters described in Bishop (2016) are other methods for computing observation increments. In the first case, the ensemble prior for an observation is approximated by a gamma distribution and observations with inverse gamma error distributions are assimilated; in the second case, the prior is inverse gamma and the observation error gamma. Again, standard linear regression is used to compute state variable increments given observation increments.

The quadratic filter of Hodyss (2012) and Hodyss et al. (2017) extends the ensemble Kalman filter by also estimating the third moment of the distributions. This requires defining auxiliary state estimation problems that assimilate pseudo-observations. Again, the increments...
for a state variable and pseudostate variables are computed using linear regression.

It is also possible to use algorithmic extensions in which a method other than linear regression is used to compute state variable increments from observation increments. For example, versions of the gamma class to compute state variable increments from observation which a method other than linear regression is used computed using linear regression.

The time mean of the ensemble mean posterior root-mean-square error is the metric used for comparing the performance of different assimilation algorithms. Unless noted, the term RMSE refers to this quantity:

\[
\text{RMSE} = \frac{1}{5000} \sum_{i=1}^{5000} \left[ \sum_{j=1}^{40} \left( x_{i,j}^T - x_{i,j}^\text{true} \right)^2 / 40 \right]^{1/2}
\]

Here, \( x \) is a state variable, the first subscript \( i \) indexes the state variable (40 here), the second subscript \( j \) indexes the time (all experiments described below will have 5000 assimilation times), superscript \( T \) refers to the true state trajectory, superscript \( u \) refers to the posterior (or updated) value, and an overbar is the ensemble mean.

Two different configurations of 40 "observation stations" are used. The "uniform" observation configuration observes at each of the 40 state variable locations. The "random" configuration observes at locations generated by 40 draws from a uniform [0, 1] distribution. When computing forward operators, the value of the state at each location is a linear interpolation between the two closest state variables. The 40 random station locations are the same as used in Anderson (2016). These locations can be computed using the random location generation procedure for the Lorenz-96 model in the default version of the DART Manhattan release (https://www.image.ucar.edu/DAReS/DART/).

Let \( x_o \) be the value of the state at an observation station. Four different forward operators are studied: identity \( y_o = x_o \); square root \( y_o = \text{sgn}(x_o)(|x_o|)^{1/2} \); cube \( y_o = x_o^3 \); and square \( y_o = x_o^2 \).

Ten initial conditions for true model trajectories were generated by starting an L96 integration from a state of all zeroes except for a 1 for the first state variable. The model was advanced for 10 million steps, and the state was saved every 1 million steps.

Five truth integrations of L96 with different assimilation periods were initiated from each of the 10 initial conditions. Assimilation periods were 1, 2, 3, 6, and 12 model time steps, and each truth integration included 5500 observation times (e.g., the truth run with period 1 had a total of 5500 time steps; the truth run for period 12 had a total of 12 x 5500).

For each of the four forward operators and for both uniform and random station configurations, several observation error variances were explored for each assimilation period; Table 1 lists the error variances. For the random station configuration, the smallest error variance with observations taken every time step for a given forward operator was selected so that the time mean prior RMSE for the best assimilations was approximately 10% of the climatological RMSE. The largest error variance for a given forward operator with assimilation every 12 model time steps has a prior RMSE that was approximately the same as the climatological RMSE.

A "case" is defined by the station configuration (uniform or random), the forward operator (identity, square root, cube, or square), the assimilation period (1, 2, 3, 6, or 12 time steps), and the observation error variance (Table 1). Six different data assimilation "treatments" are applied to each case. The treatments
use standard regression or rank regression, with the observation increments being computed by the EAKF, the perturbed observation EnKF, or the non-Gaussian RHF with the default parameter settings from the DART Manhattan release. For example, one treatment uses rank regression with the RHF observation update.

Although adaptive inflation algorithms (Zhang et al. 2004; Anderson 2009; Whitaker and Hamill 2012; El Gharamti 2018) are effective, they make the complex problem of understanding the results even more complicated. To avoid this situation, empirically tuned fixed prior multiplicative inflation was used. Tuning of each case–treatment pair was done by trying 49 pairs of inflation and Gaspari–Cohn (1999) localization half-widths; the values are listed in Table 2.

Data assimilation experiments began with each state variable of each ensemble being the true initial condition plus a draw from a unit normal. A total of 5500 data assimilation steps were done, and the first 500 steps were discarded. Five thousand steps were chosen for evaluation because it resulted in high statistical significance for detecting differences between the RMSE resulting from different treatments for selected cases (see discussion of paired $t$ tests below). Each case had 294 assimilations, 49 inflation–localization pairs for each of the six treatments.

Initially, assimilation experiments had 80 ensemble members and were only done for the first of the 10 true trajectories. For each case, the assimilation experiment using standard regression that had the lowest RMSE was identified by its observation space increment method and rank treatments for a given case. For some experiments, the assimilation failed by either crashing or producing machine infinity results. Any experiment that failed was assigned infinite RMSE.

### Table 1. Observation error variances for assimilation cases. Ensembles of 80 members were used for cases in regular font, and ensembles of 20, 40, 80, and 160 members were used for cases in italics.

<table>
<thead>
<tr>
<th>Forward operator</th>
<th>Uniform stations</th>
<th>Random stations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity</td>
<td>1 or 16</td>
<td>1, 2, 4, or 16</td>
</tr>
<tr>
<td>Square root</td>
<td>0.5 or 2</td>
<td>0.25, 0.5, 1, or 2</td>
</tr>
<tr>
<td>Cube</td>
<td>128 or 1024</td>
<td>128, 1024, or 2048</td>
</tr>
<tr>
<td>Square</td>
<td>16, 64, or 256</td>
<td></td>
</tr>
</tbody>
</table>

To estimate the statistical significance of the differences between the best standard and rank regression treatments for a given case, assimilation experiments were run for the remaining nine true trajectories for both the best rank regression and standard regression treatments. The result was ten 5000-step assimilation experiments, each with different initial conditions, for the best rank and standard treatments for each case. Paired $t$ tests for the 10 pairs of RMSE were used to provide an estimate of the significance of the differences between the best rank regression and standard regression treatments.

The experiments were also carried out with ensemble sizes of 5, 10, 20, 40, and 160 members for all cases with random station locations and identity or square root forward operators. The same procedure as for 80 members was employed. The rank regression algorithm performed very poorly with 5 and 10 members, and those results are not discussed further.

### 3. Rank regression for ensemble filters

The regression step of a standard ensemble filter computes increments for a state variable ensemble given a prior ensemble estimate of an observation and corresponding observation increments. Let $\mathbf{x}$ be the prior state ensemble, $\mathbf{y}$ be the prior observation ensemble, $\Delta\mathbf{y}$ be the observation increments, and $\Delta\mathbf{x}$ be the state increments. A subscript $n$ refers to the individual ensemble members from 1 to the ensemble size $N$. The standard regression algorithm finds the slope $b$ of the least squares fit line for the state $\mathbf{x}$ as a function of the observation $\mathbf{y}$; that is, $b$ minimizes

$$
\sum_{n=1}^{N} [x_n - (a + by_n)]^2.
$$

Then the increments for $\mathbf{x}$ are $\Delta x_n = b\Delta y_n$.

The rank regression method assumes that there is a monotonic functional relation between the observation $\mathbf{y}$ and the state $\mathbf{x}$. The implications are discussed more in the next section. The algorithm proceeds as follows (note that $\tilde{r}$ refers to a generalized rank, a real value, whereas $r$ without a tilde refers to the traditional rank, which is an integer from 1 to $N$):

1) Compute the rank statistics $r^{\text{op}} = (r^{\text{op}}_1, r^{\text{op}}_2, \ldots, r^{\text{op}}_N)$ and $r^{\text{op}} = (r^{\text{op}}_1, r^{\text{op}}_2, \ldots, r^{\text{op}}_N)$ of the prior observation ensemble $\mathbf{y}$ and prior state ensemble $\mathbf{x}$. 

### Table 2. Values of state space prior inflation and localization half-widths

<table>
<thead>
<tr>
<th>Inflation</th>
<th>Localization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>No localization</td>
</tr>
<tr>
<td>1.02</td>
<td>No localization</td>
</tr>
<tr>
<td>1.04</td>
<td>No localization</td>
</tr>
<tr>
<td>1.08</td>
<td>No localization</td>
</tr>
<tr>
<td>1.16</td>
<td>No localization</td>
</tr>
<tr>
<td>1.32</td>
<td>No localization</td>
</tr>
<tr>
<td>1.64</td>
<td>No localization</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Half-widths</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.125</td>
</tr>
<tr>
<td>0.15</td>
</tr>
<tr>
<td>0.175</td>
</tr>
<tr>
<td>0.2</td>
</tr>
<tr>
<td>0.25</td>
</tr>
<tr>
<td>0.4</td>
</tr>
</tbody>
</table>

(continued)
2) Compute the generalized rank of each member of
the ensemble of posterior (updated) observations
using the function $f$ described in the appendix:
$$r_n^u = f(y_n + \Delta y_n; y^p).$$
3) Compute the observation generalized rank incre-
ments for each member, $\Delta r_n^p = r_n^u - r_n^p$.
4) Compute the slope $b_r$ of the least squares fit line for
the prior state rank statistics as a function of the prior
observation rank statistics, that is, the value of $b_r$ that
minimizes
$$\sum_{n=1}^{N} [r_n^u - (a_r + b_r r_n^p)]^2.$$ 
5) Compute posterior state generalized rank statistics,
$$r_n^u = r_n^p + b_r \Delta r_n^p.$$ 
6) Convert the ensemble of posterior state general-
ized rank statistics to posterior state values,
$$x_n^u = f^{-1}(r_n^u, x^p).$$ 

The subscript $n$ spans the ensemble members, $n = 1, \ldots, N$.

Figure 1 displays the prior bivariate ensemble for an
observation and a state variable along with posterior
ensembles from standard and rank regression. These
examples are from the case with uniform observation
locations, a square root forward operator, assimilation
period of 12 time steps, and observation error variance
of 0.5 with the best inflation and localization for the
standard regression. In Fig. 1a, the state variable and
observation are at the same location so that the obser-
vation $y_o$ is related to the state variable $x$ by the signed
square root, $y_o = \text{sgn}(x)(|x|)^{1/2}$. The prior ensemble
shows the monotonic functional relation between the
observation and the state. The rank regression posterior
ensemble has a distribution that is similar to a subsection
of the prior ensemble. However, the standard regression
posterior from the standard regression is quite different
from any subset of the prior.

Figure 1b shows a more nuanced result. In this case,
the state variable is located two grid intervals away from
the observation so there is no functional relationship
between the observation and the state variable. Poste-
ier ensembles from the standard and rank regression
are different, but it is not obvious that one is superior to
the other.

In the uniform observation configuration, only 1 in 40
of the regressions are for a collocated observation and
state variable for which a functional relationship exists;
in the random station configuration, none of the re-
gressions have a functional relationship. It is unclear
whether the advantages of the rank regression shown
in Fig. 1a extend to situations like that in Fig. 1b. Of
course, most real assimilation problems never have
observations of a model state variable, so the uniform
observation configuration is less relevant to atmospheric
applications.

For linear, Gaussian problems, standard regression
provides the best linear, unbiased estimate (Anderson
2003). However, for general nonlinear, non-Gaussian
problems there is no form of regression that leads to a
solution that is consistent with the optimal Bayesian
solution. Rank regression is just one possible choice
and empirical validation in the next section can be used
to compare various choices.

The behavior of an ensemble filter with rank regres-
sion could be examined for specific cases for which the

\[ \text{FIG. 1. Comparison of posterior ensemble from rank regression (red plus signs) and standard regression (blue plus signs) for updates of a state variable given the same prior ensemble (open gray circles) and observation increments. The examples are taken from the case with 40 uniformly located observation stations with observational error variance of 0.5 and assimilation period of 12 with a square root forward operator. Results are shown for the case (a) in which the observation is the square root of the state variable being updated and (b) in which the state variable is located two grid intervals away from the observed variable.} \]
analytic solution is known, for instance cases in which the prior is drawn from a bivariate Gaussian with known parameters. The rank regression method does not converge to the correct solution as the ensemble size increases and has very complex behavior as ensemble size or prior correlation varies. An examination of this behavior is beyond the scope of this report.

4. Results

Initially, results for random observation configurations with 80 member ensembles are discussed for each of the different forward operators. Other ensemble sizes and uniform observation station configurations are examined subsequently.

a. Identity forward operator

Figure 2 shows the RMSE of the best treatment for each identity forward operator case with the random observation configuration. A standard regression treatment produces the lowest RMSE for all of these cases. In general, the difference in RMSE between the best rank regression treatment and the best standard regression treatment has high statistical significance (significance level \( p \) value, \( 0.000 \ 01 \) for all but three cases). For the cases with small RMSE, the best treatment generally uses an EnKF (7 of 20 cases), but for larger RMSE cases the RHF produces the lowest RMSE.

Figure 3 shows the lowest RMSE for each of the six treatments for cases with observation error variance of 4. For all five assimilation periods, the RMSE of the best standard regression treatment is a few percent smaller than that of the best rank regression. The differences are all highly significant; therefore, standard regression is clearly the treatment of choice. The observation space treatment is also important; the RHF has smaller RMSE for all standard regression cases.

Figure 4 displays the time mean of the ratio of posterior ensemble spread to RMSE for all 20 cases with identity forward operators and random observation configuration. This ratio is expected to have a time mean value of 1 for linear, Gaussian systems for which the Kalman filter solution is optimal. Figure 4a shows that the ratio varies between about 0.93 and 1.13 for rank regression treatments and is generally larger for smaller observation error variance cases. Figure 4b shows that the ratio varies between about 0.9 and 1.08 for standard regression treatments. The ratio still tends to be larger for smaller observation error variance although the relationship is not as strong as for rank regression. Minimum RMSE is not necessarily found for treatments with spread to RMSE ratios of 1, even in the relatively simple L96 system. When tuning an ensemble filter system in larger models, it is likely that trade-offs will be required between smaller RMSE and an ability to interpret the spread as an estimate of uncertainty; optimal RMSE may not be associated with good calibration of ensemble variance. These results suggest that this issue might be somewhat more important for rank regression than standard regression.

b. Square root forward operator

Figure 5 shows the RMSE of the best treatment for each of the cases for the square root forward operator. For cases with larger RMSE, the best treatments apply rank regression with an RHF. For smaller RMSE, the best treatments use standard regression with an EnKF. The differences between the best rank and standard treatments are highly significant for most of the cases in which the rank results are better; however, most of the cases in which the standard regression is better do not have statistically significant differences. In all of these cases, the median value of RMSE over the 10 experiments is smaller for rank regression with the RHF than for any of the standard regression treatments. However, for the cases with shorter observation period, a few of the rank regression experiments have much larger RMSE due to periods when the ensemble diverges from
the truth. Such periods also occur for the standard regression, but less frequently and with shorter duration. It is possible that an adaptive inflation algorithm might reduce the frequency of divergence, but that has not been explored here.

Figure 6 shows RMSE for all six treatments for cases with observation error variance of 2. Again, the differences between the best rank and standard regression treatments are normally a few percent although the difference is larger than 10% for observations every three time steps. Rank regression treatments without the RHF produce much larger RMSE than the other four treatments. Since the observation space RHF deals more effectively with non-Gaussian marginal distributions for observations than the EnKF or EAKF, this result suggests that the rank regression produces observation space priors that are significantly non-Gaussian giving the RHF a significant advantage.

Figure 7 displays the localization values for the best treatments for all 20 cases. The localization for the best rank treatments (Fig. 7a) is almost always tighter than for the corresponding standard regression treatments (Fig. 7b). The best localization generally gets tighter as the assimilation period gets longer or the observation error variance gets larger. In other words, tighter localization is required for cases with larger RMSE, however there are some exceptions. Rank regression cases with the smallest error variance (0.25) and more frequent observations (every 1, 2, or 3 time steps) have tighter localization than cases with error variance 0.5. Examination of the individual assimilation experiments shows that in cases with error variance 0.25, the ensemble occasionally diverges from the true trajectory for short periods of time. Rank regression treatments with broader localization have smaller ensemble mean errors for most of the 5000 steps, but diverge more frequently.

Figure 8 shows the inflation values for the best treatments for all 20 cases. The best standard regression treatments (Fig. 8b), often have no inflation (1.0) and never have inflation larger than 1.05. Inflation is larger
for the best rank treatments with values for assimilation periods of 2, 3, and 6 time steps that are greater than 1.05 for some observation error variances. These larger inflation values are apparently necessary to avoid periodic filter divergence that occurs in many of these cases with smaller inflation as noted above. An RHF often requires larger inflation than an EnKF or EAKF (Anderson 2010), so the larger inflation may be expected.

c. Cube forward operator

Figure 9 shows the RMSE of the best treatment for each of the cases for the cube forward operator. For cases with larger assimilation period, the best treatments use rank regression with an RHF. Smaller RMSE cases generally use standard regression and an EnKF. Several cases with intermediate RMSE use rank regression with an EnKF or standard regression with an RHF, but RMSE differences in these cases are not statistically significant.

Figure 10 shows the RMSE for all six treatments for cases with observation error variance of 1024. The difference in RMSE between the best rank and standard regression treatments is more than 10% for all assimilation periods; rank regression is best for observations taken every 6 and 12 time steps, and standard regression is better for shorter assimilation periods. Despite the large difference between the best standard and rank treatments for observations every 1, 2, and 3 steps, the differences are not statistically significant with only 10 experiments. This is because experiments are characterized by occasional filter divergence that is usually followed by a “reacquisition” of the true trajectory. In almost all cases in the entire study where the \( p \) value is less than 0.00001, all 10 experiments from the better regression have smaller RMSE than the corresponding experiment with the other regression. For the cube forward operator cases without significant differences, this is not true. The number of experiments for which standard regression has smaller RMSE than rank for these cases is between 4 and 7, but one regression may diverge much more frequently than the other. Again, these divergence issues might be reduced with the use of an adaptive inflation algorithm.

d. Square forward operator

Figure 11 shows the RMSE of the best treatment for each of the cases for the square forward operator. The function that maps state to observation no longer has an inverse in this case. The implicit assumption in the rank regression of a monotonic functional relation between the state and observation is violated in this case. The results show that standard regression with an RHF is most frequently the best treatment. There are also cases with standard regression and the EnKF for smaller
RMSE and a single instance where an EAKF with standard regression is best. The results suggest that rank regression is not generally effective for this forward operator. However, the performance of the standard regression is also probably poor in these cases relative to more general methods like particle filters (van Leeuwen 2009) or localized particle filters (Poterjoy 2016).

e. Different ensemble sizes

Ensemble sizes of 20, 40, and 160 were also used with the random observation station configuration cases with identity and square root forward operators. For identity forward operators, standard regression still produced the lowest RMSE for all cases for all ensemble sizes. However, for 20 member ensembles, the EAKF had the lowest RMSE for all but one pair of observation error variance and assimilation period. For 40 members, the EAKF, EnKF and RHF each produced the lowest RMSE for several pairs of observation error variance and assimilation period. For 160 members, the RHF produced lowest RMSE for all but four pairs of observation error variance and assimilation period, the ones for the smallest observational error variance.

Results for the square root forward operator showed more variability as a function of ensemble size. With 20 ensemble members, standard regression with the EAKF, EnKF and RHF each produced the lowest RMSE for several pairs of observation error variance and assimilation period. For 160 members, the RHF produced lowest RMSE for all but four pairs of observation error variance and assimilation period, the ones for the smallest observational error variance.

Results for the square root forward operator showed more variability as a function of ensemble size. With 20 ensemble members, standard regression with the EAKF, EnKF and RHF each produced the lowest RMSE for several pairs of observation error variance and assimilation period. For 160 members, the RHF produced lowest RMSE for all but four pairs of observation error variance and assimilation period, the ones for the smallest observational error variance.

Results for the square root forward operator showed more variability as a function of ensemble size. With 20 ensemble members, standard regression with the EAKF, EnKF and RHF each produced the lowest RMSE for several pairs of observation error variance and assimilation period. For 160 members, the RHF produced lowest RMSE for all but four pairs of observation error variance and assimilation period, the ones for the smallest observational error variance.
intermediate RMSE cases, the best treatment uses standard regression with the EAKF. For 160 members, Fig. 13 shows that results are similar to the 80 member results in Fig. 5. There are fewer cases for which the RMSE difference has the highest level of statistical significance. These results suggest that rank regression is not effective in small ensembles, but has consistent advantages for nonlinear forward operators with larger ensembles and larger RMSE. If the RMSE is small, the prior ensemble bivariate distributions of an observation and a state variable may still appear quite linear because of the limited range of uncertainty. In these cases, there is little to be gained by attempting to more accurately represent the nonlinear relation that may exist.

Figure 14 summarizes the RMSE as a function of ensemble size for the square root forward operator with assimilation period 3. In all cases, the RMSE decreases as a function of ensemble size for each observation error variance. For the smallest error variance case, 0.25, rank regression has much larger errors than standard regression and improves little with increasing ensemble size. For larger error variance, and hence larger RMSE, the improvement with ensemble size for rank regression is much larger. For observation error variance 1 and 2, the rank regression RMSE for 80 members is smaller than the 160 member RMSE for standard regression.

f. Uniform observation station configurations

Results were also produced with 80 ensemble members for uniform observation station configurations for identity, square root, and cube forward operators. In these cases, each observation impacts one state variable where the prior ensemble is exactly on the continuous curve corresponding to the forward operator. For example, Fig. 1a shows the prior ensemble from a case with the square root forward operator. Rank regression that respects this prior relation might have a distinct advantage over standard regression in this situation. However, each observation also impacts up to 39 other state variables (depending on the localization) for which the prior ensemble is more similar to the example in Fig. 1b. Results for uniform observation station configurations were generally similar to those for random station configurations. For the identity forward operator, standard regression was always better than rank regression with the RHF being used in the best treatment for larger RMSE and the EnKF for smaller RMSE. For nonlinear forward operators, rank regression with the RHF was best for larger RMSE while standard regression was better for smaller RMSE. However, the rank regression was better for smaller values of RMSE with the uniform configurations than for random configurations. This is consistent with rank regression being especially relevant for the situation in Fig. 1a.
5. Computational scaling and cost

Rank regression is considerably more complicated to compute than standard regression. Here, the discussion of cost and scaling is for the regression step for a single observation, state variable pair. For the standard regression, the cost is $O(N)$, where $N$ is the ensemble size and involves the computation of the sample covariance matrix.

Two additional computationally intensive steps are required to compute the rank regression. First, the prior state and observation ensembles must be sorted in order to compute ranks. Sorting algorithms that are $O(N \log N)$ on average and in the worst case are available. Second, given a prior ensemble, it is necessary to compute the continuous ranks for each element of a vector of $N$ values. For each of the $N$ values, this requires a search in a sorted vector of $N$ values. Using a binary search, the average and worst case are $O(\log N)$, and therefore the cost for $N$ searches is also $O(N \log N)$.

However, it is possible to reduce the computational scaling in practice. Each state variable is updated once by each observation at which point the state ensemble must be sorted. For the application here, and more generally for large geophysical applications, a single observation usually does not randomly scramble the state ensemble. Instead, an observation tends to make minor modifications so that the posterior ensemble has nearly the same rank statistics as the prior. In this case, it is possible to apply sorting algorithms that have better scaling for sorting arrays of values that are already nearly sorted. A simple insertion sort can produce $O(N)$ scaling for nearly sorted arrays. Examination of scaling results for sample cases with ensemble sizes ranging from 20 to 5120 (increasing size by factors of 2) indicates that the scaling for using an $O(N \log N)$ algorithm such as “quicksort” is as expected, but insertion sort results in $O(N)$ scaling.

A similar approach can be taken for the conversion of state values to continuous ranks given an ensemble. The values being converted are the posterior values corresponding to the prior ensemble. Again, it is normally the case that the rank of the posterior does not change drastically from its prior rank after assimilating a single observation. A linear search starting in the vicinity of the prior rank can be faster than a binary search in this case. Again, tests covering the large range of ensemble sizes indicates that the scaling is $O(N)$ for the linear search.

While this scaling is encouraging, the actual cost for the Lorenz96 cases here is significantly larger for the rank regression than for the simple regression. For just...
the regression part of the algorithm, the ratios of computational cost on a single processor are approximately 5, 8, 10, and 12 for ensemble sizes of 20, 40, 80, and 160. The regression is just a fraction of the total cost, so the ratios for the total filtering time are smaller and vary widely depending on the localization and assimilation period. For long assimilation periods with small localization, costs for the total solution can be roughly comparable.

For large geophysical applications, the relative cost of the model advance to the filtering can be much larger than for these Lorenz96 cases. Of course, this depends on details like the number of observations, the number of state variables, and the localization. Testing will be required to determine the relative cost of rank regression for large applications. In order for rank regression to be of interest, the RMSE for rank regression must be less than the RMSE for an equally costly standard regression. This requires rank regression with ensemble size $N$ to produce smaller RMSE than standard regression with an ensemble size greater than $N$. Results in Fig. 14 show that this occurs in low-order applications. Further exploration is required to see whether it is the case for some large applications.

6. Discussion and conclusions

A method for nonlinear regression in an ensemble filter has been described and tested in a low-order model. Rank regression produces smaller posterior RMSE than standard linear regression in some cases with nonlinear forward operators. One could also apply more specific forms of nonlinear regression, for instance polynomial regressions, that were tailored to the details of a specific forward operator. Because it implicitly assumes a monotonic functional relation between the state variable and observation, rank regression is less effective for forward operators that are not invertible like the square forward operator. More general filtering methods like particle filters are required for good performance in such cases.

Many other filtering methods that work well with nonlinear problems have been developed, for instance the maximum likelihood ensemble Filter (Zupanski 2005). Another example, the iterative EnKF (IEnKF) that is described in Sakov et al. (2018), is likely to perform better than any regression-based method for strongly nonlinear problems, exactly the ones for which the rank regression has advantages over standard regression. Comparison of both the quality and cost of the IEnKF and regression-based methods would be an interesting area for further research.

Simon and Bertino (2009) applied a Gaussian ana- morphosis method designed to make prior distributions more Gaussian via a univariate mapping of the prior. This leads to improved assimilation results in

Fig. 13. Time mean posterior RMSE for the best assimilation method for each of the 160 cases with randomly located observation stations and square root forward operators. The four symbols for each assimilation period are for the four values of observation error variance. A red symbol indicates that rank regression had the smallest RMSE, and a blue symbol means standard regression was best. A triangle indicates that the lowest RMSE happened with a rank histogram filter in observation space, and a circle indicates that the EnKF was best. A large symbol indicates that the difference between the best rank regression case and the best standard regression case had a $p$ value of less than 0.00001, an intermediate symbol indicates $p$ value of less than 0.01, and a small symbol is for a $p$ value of greater than 0.01.

Fig. 14. Ratio of the time mean posterior RMSE to the best RMSE with ensemble size 160 for square root observations with assimilation period 3. The first four bars for each observation error variance are for rank regression with ensemble sizes of 20, 40, 80, and 160; the second four bars are for standard regression. Bars with the same fill color are for the same ensemble size.
applications with non-Gaussian priors. One could envision doing similar remappings to make the prior relation between the state and observation more linear, as opposed to making one of the prior marginals more Gaussian.

This study also explored the performance of three different updates in observation space. The choice of observation space update often had significant impacts on the RMSE. The best observation update depended on details of the assimilation problem including the ensemble size, the assimilation period, the observation error variance, and the type of regression employed. The non-Gaussian rank histogram filter was usually essential for good performance with rank regression, suggesting that this regression led to less-Gaussian priors in observation space. The RHF also was the best choice for many problems with larger ensemble size with standard regression. The perturbed observation ensemble Kalman filter was the best choice for larger ensembles in cases with relatively small RMSE. The deterministic ensemble adjustment Kalman filter was best for many cases with small ensemble size. These results have been reported previously, but this study makes a comprehensive examination over a large number of cases.

A common theme from many comparative studies of nonlinear and non-Gaussian ensemble filter methods is that conditional application of more advanced methods is worth exploring. For instance, one could determine a priori, or try to detect on the fly, cases in which priors are particularly nonlinear or non-Gaussian and apply an appropriate method in these cases. More advanced algorithms tend to be more expensive, so conditional application may result in improved performance without excessive cost.

Even small details of standard ensemble algorithm implementations can lead to large changes in RMSE. For example, the EnKF algorithm used here sorts the posterior in observation space (Anderson 2003) so that the average value of the observation increments is minimized. This does not change the answer in linear Gaussian cases, but by minimizing the average increment in state space, it minimizes errors in cases in which the prior distributions are nonlinear. All cases for identity forward operators with random observation station configuration were repeated with an EnKF without sorting. Figure 15 shows the ratio of posterior RMSE for the best inflation/localization pair without sorting to the best case with sorting. For all cases sorting leads to smaller RMSE with a reduction of several percent in most cases. Getting good performance from an EnKF requires this sorting, which may be difficult to implement in filters that do not serially assimilate observations.

Another detail that was not explored in this study is the impact of parallel filter algorithms. The standard DART parallel algorithm (Anderson and Collins 2007) that computes forward operators for all observations at a given time at once before assimilating any of them was used here. An algorithm in which the first observation is assimilated before computing the next forward operator could significantly impact RMSE but would also impact the scalability.

Significant impacts on RMSE can be obtained by modifying ensemble filter algorithms in both observation space and for regression. Careful tuning, or good adaptive algorithms, for inflation and localization are also essential for good performance and values may vary widely for different observing systems. Low-order model results here suggest that there is lots of room to improve ensemble filter performance for larger applications. In particular, careful exploration of ensemble filter methods and tuning might lead to a reduction of RMSE for numerical weather prediction applications by many percent, improvements that are highly significant for these systems.

Acknowledgments. The author is grateful for support from the DART team, Nancy Collins, Moha El Gharamti, Johnny Hendricks, Tim Hoar, and Kevin Raeder. Stephen Anderson helped to improve and verify the presentation of the method. Thanks are given to three anonymous reviewers for comments that led to a much better presentation of the results. The National
Center for Atmospheric Research is sponsored by the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

APPENDIX

Mapping Values to Ranks and Ranks to Values

The rank regression method requires an invertible function:

\[ \hat{r} = f(\hat{x}; x) \quad \text{and} \quad \hat{x} = f^{-1}(\hat{r}; x), \quad \hat{x} = f^{-1}(\hat{r}; x), \quad (A1) \]

where \( x = (x_1, x_2, \ldots, x_N)\) is an ensemble, \( \hat{x} \) is a value, and \( \hat{r} \) is a continuous generalized rank, a real number. Let \( s = (s_1, s_2, \ldots, s_N) \) be the vector of the sorted values of \( x \) so that \( s_1 \) is the smallest value in \( x \), \( s_2 \) is the second smallest, and so on. The function \( f \) is defined as

\[ f(\hat{x}; x) = \begin{cases} 1 - b(s_1 - \hat{x}), & \hat{x} < s_1 \\ n + (\hat{x} - s_n)(s_{n+1} - s_n), & s_n \leq \hat{x} \leq s_{n+1} \\ N + b(\hat{x} - s_N), & \hat{x} > s_N \end{cases} \]

This is a linear interpolation in rank when \( \hat{x} \) is between the smallest and largest values of \( x \). An extrapolation is required if \( \hat{x} \) is smaller than the smallest value in \( x \) or larger than the largest value. Define \( r_n = n, n = 1, 2, \ldots, N \) and let \( b \) be the slope of the least squares fit line for \( r \) as a function of \( s \), that is, the value of \( b \) that minimizes

\[ \sum_{n=1}^{N} [r_n - (a + bs_n)]^2. \]

Figure A1 shows a sample of the function \( f \) for a small ensemble randomly drawn from a normal distribution.

The experiments reported here computed nearly 250 billion regressions, and there was never an instance in which an ensemble \( x \) contained duplicate values to machine precision. However, for other types of models this situation could occur frequently. In that case, \( f \) would have to be extended to handle cases in which \( s_n = s_{n+1} \).

REFERENCES


