NOTES AND CORRESPONDENCE

An Explanation for the Diagonally Predominant Property of the Positive Symmetric Ensemble Transform Matrix

Le DUC

*Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan
Meteorological Research Institute, Tsukuba, Japan*

Kazuo SAITO

*Japan Meteorological Business Support Center, Tokyo, Japan
Atmosphere and Ocean Research Institute, University of Tokyo, Kashiwa, Japan
Meteorological Research Institute, Tsukuba, Japan*

and

Daisuke HOTTA

Meteorological Research Institute, Tsukuba, Japan

(Manuscript received 8 August 2019, in final form 6 January 2020)

Abstract

In the ensemble transform Kalman filter (ETKF), an ensemble transform matrix (ETM) is a matrix that maps background perturbations to analysis perturbations. All valid ETMs are shown to be the square roots of the analysis error covariance in ensemble space that preserve the analysis ensemble mean. ETKF chooses the positive symmetric square root $\mathbf{T}^s$ as its ETM, which is justified by the fact that $\mathbf{T}^s$ is the closest matrix to the identity $\mathbf{I}$ in the sense of the Frobenius norm. Besides this minimum norm property, $\mathbf{T}^s$ is observed to have the diagonally predominant property (DPP), i.e., the diagonal terms are at least an order of magnitude larger than the off-diagonal terms.

To explain the DPP, first, the minimum norm property has been proved. Although ETKF relies on this property to choose its ETM, this property has never been proved in the data assimilation literature. The extension of this proof to the scalar multiple of $\mathbf{I}$ reveals that $\mathbf{T}^s$ is a sum of a diagonal matrix $\mathbf{D}$ and a full matrix $\mathbf{P}$ whose Frobenius norms are proportional, respectively, to the mean and the standard deviation of the spectrum of $\mathbf{T}^s$. In general cases, these norms are not much different but the fact that the number of non-zero elements of $\mathbf{P}$ is the square of the ensemble size whereas that of $\mathbf{D}$ is the ensemble size causes the large difference in the orders of elements of $\mathbf{P}$ and $\mathbf{D}$. However, the DPP is only an empirical fact and not an inherently mathematical property of $\mathbf{T}^s$. There exist certain spectra of $\mathbf{T}^s$ that break the DPP but such spectra are rarely observed in practice since their occurrences require an unrealistic situation where background errors are larger than observation errors by at least two orders of magnitude in all modes in observation space.

Corresponding author: Le Duc, Yokohama Institute for Earth Sciences, 3173-25 Showa-machi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan
E-mail: leduc@jamstec.go.jp
J-stage Advance Published Date: 16 January 2020

Keywords ensemble transform Kalman filter; ensemble transform matrix; positive symmetric square root; minimum norm property; diagonally predominant property

Citation Duc, L., K. Saito, and D. Hotta, 2020: An explanation for the diagonally predominant property of the positive symmetric ensemble transform matrix. *J. Meteor. Soc. Japan*, **98**, 455–462, doi:10.2151/jmsj.2020-022.

1. Introduction

In the ensemble transform Kalman filter (ETKF) (Bishop et al. 2001; Wang et al. 2004; Ott et al. 2004), analysis perturbations are obtained by applying a linear transformation on background perturbations. Denoting analysis and background perturbations by $n \times k$ matrices $\mathbf{X}^a$ and $\mathbf{X}^b$, respectively, this transformation is represented by a right-multiplication of $\mathbf{X}^b$ by a $k \times k$ matrix $\mathbf{T}$

$$\mathbf{X}^a = \mathbf{X}^b \mathbf{T}, \quad (1)$$

where n is the size of the state vectors, and k is the ensemble size. Here each column of $\mathbf{X}^b$ represents the difference $\mathbf{x}^b - \bar{\mathbf{x}}^b$ between each forecast member $\mathbf{x}^b$ and the ensemble mean $\bar{\mathbf{x}}^b$ calculated from all forecast members. A similar definition using the analysis ensemble is applied for the columns of $\mathbf{X}^a$.

The matrix $\mathbf{T}$ is called the ensemble transform matrix (ETM), and it is formulated so that the resulting analysis error covariance $\mathbf{P}^a = \mathbf{X}^a \mathbf{X}^{aT} / (k - 1)$ obeys the Kalman filter equation for the second moment of the posterior distribution

$$\mathbf{P}^a = \mathbf{P}^b - \mathbf{P}^b \mathbf{H}^T (\mathbf{H} \mathbf{P}^b \mathbf{H}^T + \mathbf{R})^{-1} \mathbf{H} \mathbf{P}^b, \quad (2)$$

where $\mathbf{R}$ is the observation error covariance, $\mathbf{H}: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is the observation operator, which is assumed to be linear, m is the number of observations, and $\mathbf{P}^b = \mathbf{X}^b \mathbf{X}^{bT} / (k - 1)$ is the background error covariance. Using the Sherman-Morrison-Woodbury identity, $\mathbf{P}^a$ can be rewritten as

$$\mathbf{P}^a = \frac{1}{k-1} \mathbf{X}^b (\mathbf{I} + \mathbf{Y}^{bT} \mathbf{R}^{-1} \mathbf{Y}^b / (k-1))^{-1} \mathbf{X}^{bT}, \quad (3)$$

where $\mathbf{Y}^b = \mathbf{H} \mathbf{X}^b$ is the background perturbations mapped into observation space. It is easy to check that any factorization $\mathbf{S} \mathbf{S}^T$ of the matrix $\tilde{\mathbf{P}}^a = (\mathbf{I} + \mathbf{Y}^{bT} \mathbf{R}^{-1} \mathbf{Y}^b / (k-1))^{-1}$ yields an ETM $\mathbf{S}$. Here, we call such matrix $\mathbf{S}$ a square root of $\tilde{\mathbf{P}}^a$ with a certain abuse of the mathematical notion of square root.

Suppose that the spectral decomposition of $\mathbf{Y}^{bT} \mathbf{R}^{-1} \mathbf{Y}^b / (k-1)$ is given by $\mathbf{C} \mathbf{\Gamma} \mathbf{C}^T$, where the orthogonal matrix $\mathbf{C}$ contains the eigenvectors in its

columns, and the diagonal matrix $\mathbf{\Gamma}$ contains the eigenvalues γ_i on its diagonal. Since $\mathbf{C}$ is orthogonal, we have $\mathbf{I} = \mathbf{C} \mathbf{C}^T$, and this helps to simplify $\tilde{\mathbf{P}}^a$

$$\begin{aligned} \tilde{\mathbf{P}}^a &= (\mathbf{I} + \mathbf{Y}^{bT} \mathbf{R}^{-1} \mathbf{Y}^b / (k-1))^{-1} \\ &= (\mathbf{C} \mathbf{C}^T + \mathbf{C} \mathbf{\Gamma} \mathbf{C}^T)^{-1} = \mathbf{C} (\mathbf{I} + \mathbf{\Gamma})^{-1} \mathbf{C}^T. \end{aligned} \quad (4)$$

Note that since γ_i are nonnegative, it is legitimate to take the inverse $(\mathbf{I} + \mathbf{\Gamma})^{-1}$. Based on this form, Bishop et al. (2001) proposed choosing the matrix $\mathbf{C} (\mathbf{I} + \mathbf{\Gamma})^{-1/2}$, which is clearly a square root of $\tilde{\mathbf{P}}^a$, as the ETM to be used in practice. However, such choice introduces a bias into the analysis ensemble mean since the sum of the resulting analysis perturbations differs from zero. This means that in addition to a square root of $\tilde{\mathbf{P}}^a$, an ETM has to preserve the analysis ensemble mean. Wang et al. (2004) pointed out that the positive symmetric square root $\mathbf{T}^s = \mathbf{C} (\mathbf{I} + \mathbf{\Gamma})^{-1/2} \mathbf{C}^T$ possesses this mean-preserving property. Living et al. (2008), and Sakov and Oke (2008) further demonstrated that all the square roots of $\tilde{\mathbf{P}}^a$ with the form $\mathbf{T}^s \mathbf{U}$ also have the mean-preserving property, provided that $\mathbf{U}$ is any orthogonal matrix that has the vector $\mathbf{1} = (1 \ 1 \ \dots \ 1)^T$ as its eigenvector. Thus, any $\mathbf{U}$ can be constructed by combining an arbitrary set of orthonormal vectors in the orthogonal complement of the vector $\mathbf{1}$ with the normalized one $\mathbf{1}/\sqrt{K}$ to form its eigenvectors.

Hunt et al. (2007) argued that three properties of $\mathbf{T}^s$ justify its use as the ETM in practice: (1) $\mathbf{T}^s$ preserves the analysis ensemble mean; (2) $\mathbf{T}^s$ depends continuously on $\tilde{\mathbf{P}}^a$; and (3) $\mathbf{T}^s$ is the closest matrix in the Frobenius norm to the identity matrix $\mathbf{I}$ (note that the Frobenius norm of a matrix $\mathbf{T}$ is defined as $\|\mathbf{T}\|_F = \sqrt{\sum t_{ij}^2}$, where t_{ij} are the elements of $\mathbf{T}$). However, it is clear that the first two conditions also hold for any ETM in the form $\mathbf{T}^s \mathbf{U}$. Thus, the third condition is the only condition that privileges the choice of $\mathbf{T}^s$. Underlying this condition is a simpler condition saying that when the impact of observations becomes negligible, $\mathbf{X}^a$ should be identical to $\mathbf{X}^b$, or equivalently $\mathbf{T}^s \mathbf{U} \rightarrow \mathbf{I}$ when $\mathbf{\Gamma} \rightarrow \mathbf{0}$. Since $\mathbf{T}^s \rightarrow \mathbf{I}$ in this case, this implies that $\mathbf{U} = \mathbf{I}$. This means that by simply choosing $\mathbf{U} = \mathbf{I}$, we maintain balance in analysis perturbations which

has already been achieved in background perturbations. Any rotation $\mathbf{U} \neq \mathbf{I}$, which is equivalent to taking linear combinations of background perturbations as analysis perturbations, will introduce certain imbalance in analysis perturbations. When the impact of observations becomes significant, the best we can do is to minimize the distortion of physically coherent structures in analysis perturbations introduced by assimilation through minimizing $\|\mathbf{T}^s \mathbf{U} - \mathbf{I}\|_F^2$ which also leads to $\mathbf{U} = \mathbf{I}$. This condition is of particular importance when local analyses are performed as in the local ensemble transform Kalman filter (LETKF), since we can have both grid points that are far from any observation and grid points that abound with observations nearby. It is noteworthy that Reich and Cotter (2015) placed this condition under a broader view established by the optimal transport theory to construct ETMs in the ensemble transform particle filter.

As an example, Fig. 1 shows entries of three $\mathbf{T}^s$ matrices computed at three different grid points that we obtained from an experiment with the LETKF using conventional observations and 50 ensemble members (Duc et al. 2015). A noticeable feature in this figure is the dominance of diagonal elements over off-diagonal elements (Saito et al. 2017). Reich et al. (2011) have also noticed this property of $\mathbf{T}^s$ and utilized it in a procedure called hydrostatic balancing to reduce initial imbalance in analysis ensemble members. The entries in the j th column of an ETM can be interpreted as the weights in the construction of the j th analysis perturbation from all background perturbations by a linear combination. Therefore, the dominance of diagonal elements in $\mathbf{T}^s$ implies that the largest contribution to the j th analysis perturbation comes from the j th background perturbation.

How can we explain such an interesting property of $\mathbf{T}^s$, which we call the diagonally predominant property (DPP) of $\mathbf{T}^s$? The first possible explanation can be traced back to Hunt et al. (2007) in their argument for choosing $\mathbf{T}^s$ as the ETM in LETKF. Their argument implies that the DPP of $\mathbf{T}^s$ is likely due to its similarity to the identity $\mathbf{I}$. In Fig. 1c, the average value of the diagonal elements is approximately 0.99, which seems to verify this hypothesis. However, the corresponding values in Figs. 1a and 1b are 0.39 and 0.69, respectively, which are much smaller than 1; consequently, it is difficult to see the similarity between these $\mathbf{T}^s$ matrices and the identity $\mathbf{I}$. This suggests that the distance between $\mathbf{T}^s$ and $\mathbf{I}$ should be quantified, which may give us a clue to understanding these large deviations from 1.

Hunt et al. (2007) did not provide any proof for this minimum norm property of $\mathbf{T}^s - \mathbf{I}$, but referred to Ott et al. (2002, 2004) for the proof. However, Ott et al. (2004) tried to seek a left-multiplication $n \times n$ matrix $\mathbf{Z}$ to transform $\mathbf{X}^b$ to $\mathbf{X}^a$ subjected to the constraint in Eq. (2) rather than a right-multiplication matrix $\mathbf{T}$ as in (1)

$$\mathbf{X}^a = \mathbf{Z}\mathbf{X}^b. \quad (5)$$

The matrix $\mathbf{Z}$ was chosen so that the distance between $\mathbf{X}^a$ and $\mathbf{X}^b$ was minimized. They also demonstrated that a right-multiplication matrix $\mathbf{T}$, which they called $\mathbf{Y}$, can be derived from $\mathbf{Z}$ so that $\mathbf{X}^b \mathbf{T}$ yields the same $\mathbf{X}^a$. However, the fact that $\mathbf{T}^s$ is the closest ETM to the identity $\mathbf{I}$ has not yet been proven.

In this paper, first, we fill in this theoretical gap in the literature by providing a proof for the minimum norm property of $\mathbf{T}^s - \mathbf{I}$ in Section 2. This property itself does not give much insight into the DPP. Therefore, a new mathematical view is needed to make the problem more accessible. This new mathematical treatment is presented in Section 3. Finally, Section 4 summarizes the results obtained in the paper and suggests potential applications of the DPP.

2. Minimum norm property

To prove this property, we find the ETM that minimizes the Frobenius norm of $\mathbf{T} - \mathbf{I}$ among all mean-preserving ETMs $\mathbf{T}$, which is expected to be $\mathbf{T}^s$. In fact, this is a consequence of a more general result:

Among all square roots $\mathbf{S}$ of a symmetric positive-definite matrix $\mathbf{M}$, the positive symmetric square root $\mathbf{S}^*$ is the closest matrix to the identity $\mathbf{I}$. Furthermore, the squared distance $\|\mathbf{S}^* - \mathbf{I}\|_F^2$ between $\mathbf{S}^*$ and $\mathbf{I}$ is given by

$$\|\mathbf{S}^* - \mathbf{I}\|_F^2 = \sum_i (\lambda_i - 1)^2, \quad (6)$$

where λ_i are the positive square roots of the eigenvalues of $\mathbf{M}$. The detailed proof is given in Appendix A.

When we apply this result into our case with $\mathbf{M} = \tilde{\mathbf{P}}^a$ and $\mathbf{S} = \mathbf{T}$, $\mathbf{T}^s$ will play the role of $\mathbf{S}^*$. Since $\mathbf{T}^s$ also has the mean-preserving property, this verifies the minimum norm property of $\mathbf{T}^s - \mathbf{I}$. To estimate the squared distance $\|\mathbf{T}^s - \mathbf{I}\|_F^2$ as given in (6), we need to know the spectrum of $\mathbf{T}^s$, which depends on the spectrum of $\tilde{\mathbf{P}}^a$. This, in turn, is determined by the spectrum of the matrix $\mathbf{Y}^{bT} \mathbf{R}^{-1} \mathbf{Y}^b / (k - 1)$, which we denote by $\mathbf{G}$. Therefore, (4) implies the following form of the eigenvalues of $\mathbf{T}^s$

Fig. 1. Entries of 50×50 $\mathbf{T}^s$ matrices at different grid points obtained from an LETKF experiment with conventional observations. The grid point in panel (a) is in the region with dense observations, whereas the grid point in panel (c) is far from observations. The grid point in panel (b) is between those in panels (a) and (c). From left to right, the average values of the diagonal elements are 0.39, 0.69, and 0.99.

$$\lambda_i = \frac{1}{\sqrt{1 + \gamma_i}}, \quad (7)$$

where γ_i are the eigenvalues of $\mathbf{G}$. It is easy to verify that all λ_i are bounded within $[0, 1]$.

Plugging λ_i in (7) into (6) $\|\mathbf{T}^s - \mathbf{I}\|_F^2$ has the following value

$$\|\mathbf{T}^s - \mathbf{I}\|_F^2 = \sum_{i=1}^r (\lambda_i - 1)^2 = \sum_{i=1}^r \left(\frac{1}{\sqrt{1 + \gamma_i}} - 1 \right)^2, \quad (8)$$

where r is the rank of $\mathbf{G}$. Noting that $(\lambda_i - 1)^2 \leq 1$ because $0 \leq \lambda_i \leq 1$ for all i , we see that $\|\mathbf{T}^s - \mathbf{I}\|_F^2$ is at most r . In general, when the number of influence observations m that contribute to the analysis at a given grid point is less than the number of ensemble members k , r is equal to m . Therefore, if a point is far from all observations, r will become smaller, leading to a smaller distance $\|\mathbf{T}^s - \mathbf{I}\|_F^2$. This explains why $\mathbf{T}^s$ in Fig. 1c is almost identical to $\mathbf{I}$. However, in the region with many observations, r attains its maximum $k - 1$ and all γ_i are greater than 0, which makes all λ_i deviate from 1 and $\mathbf{T}^s$ to be more dissimilar to $\mathbf{I}$, as illustrated in Fig. 1a.

3. Diagonally predominant property

The minimum norm property of $\mathbf{T}^s - \mathbf{I}$ alone cannot explain why we see the DPP of $\mathbf{T}^s$, especially when all λ_i are far from 1 and as a result, $\mathbf{T}^s$ becomes more dissimilar to $\mathbf{I}$. The choice of $\mathbf{I}$ for comparison with $\mathbf{T}^s$ is reasonable when influence of observations is small and we expect that the analysis perturbations are

more or less similar to the background perturbations. However, when influence of observations becomes significant, we expect that the observations will help us to reduce the uncertainty in the background perturbations considerably. In such a case, the identity $\mathbf{I}$ is clearly not a good choice for comparison, and we should instead use something like $\alpha\mathbf{I}$ where α is the reduction factor resulting from assimilation of observations, i.e. $0 < \alpha < 1$.

Therefore, we extend the minimization problem in Section 2 by adding a parameter α , and find the ETM $\mathbf{T}$ and the parameter α that minimize the distance $\|\mathbf{T} - \alpha\mathbf{I}\|_F$. Again, we can obtain a more general result for an arbitrary symmetric positive-definite matrix: Among all square roots $\mathbf{S}$ of a symmetric positive-definite matrix $\mathbf{M}$, the positive symmetric square root $\mathbf{S}^*$ is the closest matrix to a scalar multiple of the identity $\mathbf{I}$. The scalar multiple of $\mathbf{I}$ closest to $\mathbf{S}^*$ in this case is the matrix $\alpha^*\mathbf{I} = \bar{\lambda}\mathbf{I}$, where $\bar{\lambda}$ is the mean of the eigenvalues of $\mathbf{S}^*$. Furthermore, the distance $\|\mathbf{S}^* - \alpha^*\mathbf{I}\|_F$ between $\mathbf{S}^*$ and $\alpha^*\mathbf{I}$ is given by

$$\|\mathbf{S}^* - \alpha^*\mathbf{I}\|_F = \sqrt{k}\sigma_\lambda, \quad (9)$$

where σ_λ is the standard deviation of the eigenvalues of $\mathbf{S}^*$. The detailed proof is given in Appendix B.

Applied into our case where $\mathbf{S}^*$ becomes $\mathbf{T}^s$ and λ_i are given in (7), this result suggests that we can decompose $\mathbf{T}^s$ into the sum of a diagonal matrix $\mathbf{D}$ and a perturbation matrix $\mathbf{P}$

$$\mathbf{T}^s = \mathbf{D} + \mathbf{P} = \bar{\lambda}\mathbf{I} + \mathbf{P}. \quad (10)$$

All the off-diagonal entries of $\mathbf{T}^s$ are also the off-

Fig. 2. Histograms of elements of the perturbation matrices $\mathbf{P}$ corresponding to the ETMs $\mathbf{T}^s$ in Fig. 1. Statistics are derived from the absolute values of all elements. To have a fair comparison, these values are normalized using their means and standard deviations. In each case, nine bins with the same width set to the standard deviation are used to construct the histogram. The typical magnitudes $\sigma_\lambda/\sqrt{k}$ of the entries of $\mathbf{P}$ as suggested by the theory are denoted by “Estimated”.

diagonal entries of $\mathbf{P}$. While the Frobenius norm of $\mathbf{D}$ is $\sqrt{k}\bar{\lambda}$, the Frobenius norm of $\mathbf{P}$ can be derived from (9), which is $\sqrt{k}\sigma_\lambda$. This enables us to estimate the typical magnitude of an element in $\mathbf{P}$ as $\sqrt{\|\mathbf{P}\|/k^2} = \sigma_\lambda/\sqrt{k}$. To check the validity of this estimation, for each matrix $\mathbf{T}^s$ given in Fig. 1, we plot in Fig. 2 the histograms of the absolute values of the elements of its corresponding matrix $\mathbf{P}$. It can be seen that approximately 90 % of the elements concentrate on the 1-sigma interval around the mean. The typical magnitudes $\sigma_\lambda/\sqrt{k}$ for the entries of $\mathbf{P}$ are also given in Fig. 2 as “Estimated” for comparison. These values reflect quite well the magnitudes of elements of $\mathbf{P}$, which are also the off-diagonal elements of $\mathbf{T}^s$. By contrast, the typical magnitude of an element along the diagonal of $\mathbf{D}$ is simply $\bar{\lambda}$. This is exactly the average value of the diagonal elements of $\mathbf{T}^s$, since $\bar{\lambda} = \text{tr}(\mathbf{T}^s)/k$.

To roughly estimate the difference in the orders between the diagonal and off-diagonal elements of $\mathbf{T}^s$, we assume that in the general case without any information about $\mathbf{G}$, all λ_i are realizations of a uniform distribution over $[0, 1]$. Based on this assumption, $\bar{\lambda} = 1/2$, $\sigma_\lambda = 1/\sqrt{12}$, and the ratio $\bar{\lambda}\sqrt{k}/\sigma_\lambda$ for an ensemble size of 50 members is approximately 12. For the $\mathbf{T}^s$ in Fig. 1a which is a matrix $\mathbf{T}^s$ associated with assimilation of dense observations, this ratio is approximately 16, which is not too different from the rough estimate of 12. This means that the diagonal elements of $\mathbf{T}^s$ are generally an order of magnitude larger than the off-diagonal elements for an ensemble size on the order of 50. This difference in orders tends to become larger when all λ_i cluster around 1, like the case in Fig. 2c, with the ratio of approximately 315 when the impact of observations is small.

Although we call $\mathbf{P}$ the perturbation matrix to emphasize that $\mathbf{T}^s$ is mostly similar to the diagonal matrix

$\mathbf{D}$, its norm is generally not much smaller than the norm of $\mathbf{D}$, i.e., $\sqrt{k}\sigma_\lambda$ compared to $\sqrt{k}\bar{\lambda}$. If we again use the assumption of the uniform distribution for λ_i , the norm of $\mathbf{D}$ is only $\sqrt{3}$ times larger than that of $\mathbf{P}$. Recalling that the square of the Frobenius norm of a matrix is the sum of squares of all the elements of that matrix, we can explain why the typical magnitudes of elements of the two matrices are quite different. This sum in the case of $\mathbf{D}$ only comprises k diagonal elements, whereas that in the case of $\mathbf{P}$ comprises k^2 elements. As a result, a typical element of $\mathbf{P}$ is $\sqrt{3k}$ times less than a typical element of $\mathbf{D}$. This ratio is equivalent to one order of magnitude for a typical ensemble size $k = 100$ in practice. It is this large difference in the number of non-zero elements between $\mathbf{P}$ and $\mathbf{D}$ that causes the large difference in the orders of magnitude between the diagonal and off-diagonal elements of $\mathbf{T}^s$.

It is informative to discuss the condition or situation in which the DPP may cease to hold. In this paragraph, we aim to demonstrate that such a situation rarely happens in practice. The quantity $\bar{\lambda}\sqrt{k}/\sigma_\lambda$, which characterizes the difference in the orders between the diagonal and off-diagonal elements of $\mathbf{T}^s$, depends not only on k but also on the ratio $\bar{\lambda}/\sigma_\lambda$. Thus, it is anticipated that the diagonal predominance will no longer be observed if this ratio is of the order $1/\sqrt{k}$ or lower order. This can happen if $\bar{\lambda}/\sigma_\lambda \leq 1/\sqrt{k}$, which is equivalent to $\bar{\lambda}^2 \geq (k+1)\bar{\lambda}^2$. The fact that λ_i are bounded between 0 and 1 implies that $\bar{\lambda}^2 \leq \bar{\lambda}$. Therefore, the inequality $\bar{\lambda} \geq (k+1)\bar{\lambda}^2$ is the necessary condition so that the DPP ceases to hold. It is easy to check that this inequality gives an upper bound $1/(k+1)$ for $\bar{\lambda}$. Therefore, for diagonal and off-diagonal elements of $\mathbf{T}^s$ to be of comparable order, for a typical ensemble size $k = 100$, $\bar{\lambda}$ must be less than

10^{-2} , which is equivalent to the lower bound 10^4 for γ_i . In other words, background errors must be 100 times greater than observation errors almost in all modes, which cannot be met in practice.

4. Summary and conclusion

In this study we have developed a mathematical framework to explain the DPP observed in the positive symmetric ETM $\mathbf{T}^s$ used in ETKF. This property has a close connection with the minimum norm property of this matrix, which states that $\mathbf{T}^s$ is the closest matrix in the Frobenius norm to the identity $\mathbf{I}$ among all potential ETMs. There exist many valid ETMs that can yield the same analysis error covariance without altering the analysis ensemble mean. ETKF relies on the minimum norm property of $\mathbf{T}^s$ to justify its choice of this matrix as the ETM in its formulation. This property has been stated but has never been proved in the data assimilation literature.

Therefore, our first step in understanding why the diagonal terms dominate over the off-diagonal terms in $\mathbf{T}^s$ is to prove the minimum norm property, since this proof can provide an important guide in the subsequent mathematical argument. We have found that this property follows from an important theorem on square roots $\mathbf{S}$ of a symmetric positive-definite matrix $\mathbf{M}$: Among all square roots of $\mathbf{M}$, the positive symmetric square root $\mathbf{S}^*$ is the closest matrix to the identity $\mathbf{I}$. This theorem suggests that instead of $\mathbf{I}$, we can further compare $\mathbf{S}$ with a scalar multiple of $\mathbf{I}$. This leads to another important theorem: Among all square roots of $\mathbf{M}$, $\mathbf{S}^*$ is the closest matrix to a scalar multiple of $\mathbf{I}$, and the scalar multiple of $\mathbf{I}$ closest to $\mathbf{S}^*$ in this case is $\bar{\lambda}\mathbf{I}$, where $\bar{\lambda}$ is the average of the eigenvalues of $\mathbf{S}^*$.

This is the second theorem that gives an explanation for the DPP of $\mathbf{T}^s$. The matrix $\mathbf{T}^s$ can be decomposed into a sum of a diagonal matrix $\mathbf{D}$ and a perturbation matrix $\mathbf{P}$ which is a full matrix whose off-diagonal elements are identical to those of $\mathbf{T}^s$. The Frobenius norm of $\mathbf{D}$ is $\sqrt{k}\bar{\lambda}$, whereas that of $\mathbf{P}$ is $\sqrt{k}\sigma_\lambda$. Thus, the two norms are associated with the first two moments of the spectrum of $\mathbf{T}^s$. Although the norm of $\mathbf{P}$ is in general not much smaller than that of $\mathbf{D}$, the fact that the number of non-zero elements of $\mathbf{P}$ is proportional to the square of the ensemble size causes its elements to be much smaller than the elements of $\mathbf{D}$ whose number of non-zero elements is only the ensemble size. This explains why the diagonal elements of $\mathbf{T}^s$ dominate over its off-diagonal elements.

It must be emphasized that the DPP is not an inherently mathematical property of $\mathbf{T}^s$. There exist certain distributions of λ_i that break the DPP such as when all

λ_i cluster around zero. However, realizations of such distributions rarely occur in practice since their occurrences require very large differences in orders between background errors and observation errors at all scales. Thus, it is almost certain that we do not observe the ETMs that do not possess the DPP in practice.

A natural question is how we utilize this DPP in practice. The hydrostatic balancing procedure in Reich et al. (2011) is an example of its application. This property is a consequence of the fact that a mean-preserving ETM derived from $\tilde{\mathbf{P}}^a$ most resembles a scalar multiple of $\mathbf{I}$ when it is $\mathbf{T}^s$, and the corresponding diagonal matrix is $\bar{\lambda}\mathbf{I}$. This suggests that we can use the best approximation $\bar{\lambda}\mathbf{I}$ as a replacement for $\mathbf{T}^s$ in ETKF. In fact, the diagonal ETM $\bar{\lambda}\mathbf{I}$ can be shown to be a special case of what we call the diagonal ETKF (Duc et al. 2019; manuscript in revision).

The diagonal approximation, when applied to ensemble forecasting, is equivalent to an ensemble generation method that combines two aspects of ETKF and the breeding method (Toth and Kalnay 1993). In this method, we rescale all forecast perturbations by the same factor $\bar{\lambda}$ at the end of each breeding cycle. Compared to the original breeding method proposed in Toth and Kalnay (1997), where the rescaling factors are derived from climatological statistics, our proposed method is appealing in that it can adaptively incorporate ETKF-induced information on the posterior error variances that reflects both the flow-dependent forecast errors and the observation network. The advantage of the proposed method, in comparison to the ensemble generation by ETKF, is presumably better dynamical balance that should be achieved by not mixing up different perturbations. We note, however, that precisely for the same reason, it will also be more prone to converge all perturbations into a single fastest growing mode, losing the ETKF's merit of avoiding such a convergence (Wang and Bishop 2003). We postulate that this issue would be less problematic in applications such as regional ensemble prediction systems, in which lateral boundary perturbations are fed into ensemble members in different directions during the course of integration as demonstrated by Saito et al. (2012). We will test this potential application of $\bar{\lambda}\mathbf{I}$ as an ETKF-based ensemble generation method in our next study.

Acknowledgments

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) through the Strategic Programs for Innovative Research, the FLAGSHIP2020 project (Advancement of

meteorological and global environmental predictions utilizing observational “Big Data”, hp160229), and the JSPS Grant-in-Aid for Scientific Research (Study of optimum perturbation methods for ensemble data assimilation, 16H04054).

Appendix A: The minimum point of the distance $\|\mathbf{S} - \mathbf{I}\|_F$ where $\mathbf{S}$ is a square root of a symmetric positive-definite matrix $\mathbf{M}$

Let $\mathbf{M}$ be a $k \times k$ symmetric positive-definite matrix, and $\mathbf{S}$ be a square root of $\mathbf{M}$. Suppose that $\mathbf{S}$ has the singular value decomposition (SVD) $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$. Since $\mathbf{S}$ is a square root of $\mathbf{M}$ we have

$$\mathbf{S}\mathbf{S}^T = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^T = \mathbf{M}. \quad (\text{A1})$$

This equation indicates that all vectors $\mathbf{u}_i$ must form an eigenbasis of $\mathbf{M}$, and the squares of all scalars λ_i have to be the eigenvalues of $\mathbf{M}$. Thus, any eigen-decomposition of $\mathbf{M}$ yields the left singular vectors and the singular values of $\mathbf{S}$, and the right-singular vectors $\mathbf{v}_i$ can be chosen arbitrarily from any orthonormal basis.

Given this general form of $\mathbf{S}$, we will find $\mathbf{S}^*$ that minimizes the squared Frobenius norm of $\mathbf{S} - \mathbf{I}$. From the definition of the Frobenius norm of a matrix, we have

$$\begin{aligned} \|\mathbf{S} - \mathbf{I}\|_F^2 &= \text{tr}[(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T - \mathbf{I})(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T - \mathbf{I})^T] \\ &= \text{tr}(\mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^T) - \text{tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T) \\ &\quad - \text{tr}(\mathbf{V}\mathbf{\Lambda}\mathbf{U}^T) + \text{tr}(\mathbf{I}), \end{aligned} \quad (\text{A2})$$

where the symbol tr stands for the trace operator. Since $\mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^T = \mathbf{M}$ and the trace is invariant under the transpose, this can be simplified as

$$\|\mathbf{S} - \mathbf{I}\|_F^2 = \text{tr}(\mathbf{M}) + \text{tr}(\mathbf{I}) - 2\text{tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T). \quad (\text{A3})$$

Therefore, $\|\mathbf{S} - \mathbf{I}\|_F^2$ will attain its minimum when $\text{tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T)$ attains its maximum.

Applying the cyclic property of the trace operator, we obtain the value of $\text{tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T)$

$$\text{tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T) = \text{tr}(\mathbf{\Lambda}\mathbf{V}^T\mathbf{U}) = \sum_i \lambda_i \langle \mathbf{u}_i, \mathbf{v}_i \rangle, \quad (\text{A4})$$

where the symbol $\langle \cdot \rangle$ denotes the inner product. Then it is easy to see that

$$\sum_i \lambda_i \langle \mathbf{u}_i, \mathbf{v}_i \rangle \leq \sum_i \lambda_i. \quad (\text{A5})$$

Since all λ_i are constants that are determined by the spectrum of $\mathbf{M}$, this implies that $\text{tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T)$ attains its

maximum when $\langle \mathbf{u}_i, \mathbf{v}_i \rangle = 1$ or equivalently $\mathbf{u}_i = \mathbf{v}_i$. In other words, $\|\mathbf{S} - \mathbf{I}\|_F^2$ attains its minimum when $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, which is the positive symmetric square root of $\mathbf{M}$. Now the calculation of the squared distance $\|\mathbf{S}^* - \mathbf{I}\|_F^2$ is straightforward

$$\begin{aligned} \|\mathbf{S}^* - \mathbf{I}\|_F^2 &= \text{tr}(\mathbf{M}) + \text{tr}(\mathbf{I}) - 2\sum_i \lambda_i \\ &= \sum_i \lambda_i^2 - 2\sum_i \lambda_i + k \\ &= \sum_i (\lambda_i - 1)^2. \end{aligned} \quad (\text{A6})$$

Note that we have used the fact that $\text{tr}(\mathbf{M}) = \sum_i \lambda_i^2$ to obtain this equation.

Appendix B: The minimum point of the distance $\|\mathbf{S} - \alpha\mathbf{I}\|_F$ where $\mathbf{S}$ is a square root of a symmetric positive-definite matrix $\mathbf{M}$ and α is a scalar variable

The minimization problem here is an extension of the minimization problem in Appendix A. Now instead of the squared distance $\|\mathbf{S} - \mathbf{I}\|_F^2$, we minimize the squared distance $\|\mathbf{S} - \alpha\mathbf{I}\|_F^2$, where we augment the minimization space by introducing a new variable α in addition to the matrix $\mathbf{S}$. When the same mathematical notions and operations are used as in Appendix A, this squared distance can be represented as

$$\begin{aligned} \|\mathbf{S} - \alpha\mathbf{I}\|_F^2 &= \text{tr}(\mathbf{M}) - 2\alpha\text{tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T) + \alpha^2\text{tr}(\mathbf{I}) \\ &= \sum_i \lambda_i^2 - 2\alpha\sum_i \lambda_i \langle \mathbf{u}_i, \mathbf{v}_i \rangle + \alpha^2 k. \end{aligned} \quad (\text{B1})$$

By dividing this distance by k , and denoting $\bar{\lambda}^2 = \sum_i \lambda_i^2/k$, and $\beta = \sum_i \lambda_i \langle \mathbf{u}_i, \mathbf{v}_i \rangle/k$, we obtain a function of two variables

$$d(\alpha, \beta) = \|\mathbf{S} - \alpha\mathbf{I}\|_F^2/k = \alpha^2 - 2\alpha\beta + \bar{\lambda}^2, \quad (\text{B2})$$

where β is bounded within $[-\bar{\lambda}, \bar{\lambda}]$. The bounded interval of β can be verified from the following inequality

$$|\beta| = \left| \frac{\sum_i \lambda_i \langle \mathbf{u}_i, \mathbf{v}_i \rangle}{k} \right| \leq \frac{\sum_i \lambda_i |\langle \mathbf{u}_i, \mathbf{v}_i \rangle|}{k} \leq \frac{\sum_i \lambda_i}{k} = \bar{\lambda}. \quad (\text{B3})$$

Note that the equality occurs in two cases, when $\mathbf{u}_i = \mathbf{v}_i$ and $\mathbf{u}_i = -\mathbf{v}_i$ for all i . In other words, $\beta = \bar{\lambda}$ if $\mathbf{S}$ is the positive symmetric square root of $\mathbf{M}$, and $\beta = -\bar{\lambda}$ if $\mathbf{S}$ is the negative symmetric square root of $\mathbf{M}$.

The critical point of the function $d(\alpha, \beta)$ is (α, β)

$= (0, 0)$, which is a saddle point. This means that $d(\alpha, \beta)$ can only attain its minimum on the boundary of its domain. When β is fixed, $d(\alpha, \beta)$ is a quadratic function, and it attains its minimum when α is equal to this fixed value. Therefore, $d(\alpha, \beta)$ has two minimum points $(\bar{\lambda}, \bar{\lambda})$ and $(-\bar{\lambda}, -\bar{\lambda})$. It is easy to understand why we have two minimum points here, because the function $d(\alpha, \beta)$ is symmetric with respect to (α, β) : $d(-\alpha, -\beta) = d(\alpha, \beta)$. Thus, we only need to consider the positive minimum point in this case and this point yields the minimum point $(\mathbf{S}^*, \alpha^*) = (\mathbf{U}\mathbf{A}\mathbf{V}^T, \bar{\lambda})$ of the squared distance $\|\mathbf{S} - \alpha\mathbf{I}\|_F^2$. Again, like the minimization problem in Appendix A, we obtain the positive symmetric square root of $\mathbf{M}$ at the minimum point. Thus, we have found that $\mathbf{S}$ resembles a scalar multiple of $\mathbf{I}$ most when it is the positive symmetric square root, and the scalar multiple of $\mathbf{I}$ in this case is $\bar{\lambda}\mathbf{I}$. The minimum squared distance $\|\mathbf{S}^* - \alpha^*\mathbf{I}\|_F^2$ is easy to estimate now

$$\begin{aligned}\|\mathbf{S}^* - \alpha^*\mathbf{I}\|_F^2 &= kd(\bar{\lambda}, \bar{\lambda}) = k(\bar{\lambda}^2 - \bar{\lambda}^2) \\ &= k((\bar{\lambda} - \bar{\lambda})^2) = k\sigma_\lambda^2,\end{aligned}\quad (\text{B4})$$

where σ_λ is the standard deviation of the spectrum of $\mathbf{S}^*$.

References

- Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part 1: Theoretical aspects. *Mon. Wea. Rev.*, **129**, 420–436.
- Duc, L., T. Kuroda, K. Saito, and T. Fujita, 2015: Ensemble Kalman Filter data assimilation and storm surge experiments of tropical cyclone Nargis. *Tellus A*, **67**, 25941, doi:10.3402/tellusa.v67.25941.
- Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, **230**, 112–126.
- Livingston, D. M., S. L. Dance, and N. K. Nichols, 2008: Unbiased ensemble square root filters. *Physica D*, **237**, 1021–1028.
- Ott, E., B. R. Hunt, I. Szunyogh, M. Corazza, E. Kalnay, D. J. Patil, J. A. Yorke, A. V. Zimin, and E. J. Kostelich, 2002: Exploiting local low dimensionality of the atmospheric dynamics for efficient ensemble Kalman filtering. *arXiv.org, e-Print Archive*, 24 pp. [Available at <http://arxiv.org/abs/physics/0203058v3>.]
- Ott, E., B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. J. Patil, and J. A. Yorke, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A*, **56**, 415–428.
- Reich, S., and C. Cotter, 2015: *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge University Press, 297 pp.
- Reich, H., A. Rhodin, and C. Schraff, 2011: LETKF for the nonhydrostatic regional model COSMO-DE. *COSMO Newsl.*, **11**, 27–31.
- Saito, K., H. Seko, M. Kunii, and T. Miyoshi, 2012: Effect of lateral boundary perturbations on the breeding method and the local ensemble transform Kalman filter for mesoscale ensemble prediction. *Tellus A*, **64**, 11594, doi:10.3402/tellusa.v64i0.11594.
- Saito, K., M. Kunii, L. Duc, and T. Kurihana, 2017: Perturbation Methods for Ensemble Data Assimilation. *RIKEN International Symposium on Data Assimilation*, Kobe, Japan. [Available at http://www.data-assimilation.riken.jp/risda2017/program/abstracts/pdf/19_2_K.Saito.pdf.]
- Sakov, P., and P. R. Oke, 2008: Implications of the form of the ensemble transformation in the ensemble square root filters. *Mon. Wea. Rev.*, **136**, 1042–1053.
- Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.
- Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319.
- Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158.
- Wang, X., C. H. Bishop, and S. J. Julier, 2004: Which is better, an ensemble of positive-negative pairs or a centered spherical simplex ensemble? *Mon. Wea. Rev.*, **132**, 1590–1605.