*Article*

# Geometric Analysis of Conditional Bias-Informed Kalman Filters

Haksu Lee [1,*], Haojing Shen [2] and Dong-Jun Seo [2]

1 Len Technologies, Oak Hill, VA 20171, USA
2 Department of Civil Engineering, University of Texas at Arlington, Arlington, TX 76019, USA; haojing.shen@mavs.uta.edu (H.S.); djseo@uta.edu (D.-J.S.)
* Correspondence: haksulee@hanmail.net; Tel.: +1-240-413-1133

**Abstract:** This paper presents a comparative geometric analysis of the conditional bias (CB)-informed Kalman filter (KF) with the Kalman filter (KF) in the Euclidean space. The CB-informed KFs considered include the CB-penalized KF (CBPKF) and its ensemble extension, the CB-penalized Ensemble KF (CBEnKF). The geometric illustration for the CBPKF is given for the bi-state model, composed of an observable state and an unobservable state. The CBPKF co-minimizes the error variance and the variance of the Type-II error. As such, CBPKF-updated state error vectors are larger than the KF-updated, the latter of which is based on minimizing the error variance only. Different error vectors in the Euclidean space imply different eigenvectors and covariance ellipses in the state space. To characterize the differences in geometric attributes between the two filters, numerical experiments were carried out using the Lorenz 63 model. The results show that the CBEnKF yields more accurate confidence regions for encompassing the truth, smaller errors in the ensemble mean, and larger norms for Kalman gain and error covariance matrices than the EnKF, particularly when assimilating highly uncertain observations.

**Keywords:** geometric analysis; conditional bias; CBPKF; CBEnKF; covariance ellipse

## 1. Introduction

With highly uncertain observations and model dynamics, KF [1] and EnKF [2] estimates tend to be conditionally biased. The conditional bias (CB)-informed Kalman filter (KF) is designed to improve the estimation of extreme states in geophysical data assimilation, particularly at the onset of extreme phenomena. The CB-informed KFs developed to date include the CB-penalized Kalman filter (CBPKF; [3,4]) and its ensemble extension, the CB-penalized Ensemble Kalman filter (CBEnKF; [5]), and aim at addressing CB [3–5] to improve estimation of extreme states. With extreme precipitation, droughts, and floods occurring more frequently in many parts of the globe, the estimation and prediction of extremes is an increasingly important topic in hydrology. With the increasing availability of diverse sources of observations, hydrologic data assimilation has a large role to play in improving state estimation. The CB-informed KF is an effort to help address both.

There are two types of CB, Type I and Type II. The Type-I CB is defined as $E[X \mid X^* = x^*] - x^*$ where $X$, $X^*$, and $x^*$ represent the unknown truth, the estimate, and the realization of $X^*$, respectively. The Type-II CB is defined as $E[X^* \mid X = x] - x$ where $x$ represents the realization of $X$. The Type-I CB is associated with false alarms which can be reduced by calibration. The Type-II CB is associated with failure to detect an event and cannot be reduced by calibration. The CBPKF minimizes a linearly weighted sum of the error covariance and the expectation of the Type-II error squared. With skillful specification of the weight for the latter, the CBPKF improves estimation of extremes over the KF while slightly increasing the unconditional MSE [3,4].

Though extensively studied in statistics, econometrics, meteorology, and hydrology [6–16], CB has gained interest only recently in data assimilation. Relevant studies to date include

the CBPKF [3,4] and the CBEnKF [5]—the CBEnKF application to flood forecasting [5], adaptive filtering for the CBPKF [17] and the CBEnKF [18], and the variance-inflated KF-based approximation of the CBPKF for reduced computation [17,18]. Improving understanding of how the CB-informed KF compares with conventional data assimilation techniques, such as the KF, is crucial for interpreting and analyzing its application results, as well as identifying areas for further development. To that end, this study provides a geometric analysis of the CB-informed KF and compares performance with the EnKF via numerical experiments.

Stochastic variables may be expressed as vectors in the Euclidean domain, which can be used to geometrically illustrate vectors of innovation, observation errors, and forecast and analysis errors in the Euclidean domain [19]. The geometric interpretation of state estimation thus helps understand how Kalman gain shape to analysis error vectors and assimilation solutions for observable as well as unobservable states [19]. Error covariance matrices may also be analyzed geometrically in the state space and characterized with eigenvalues, eigenvectors, and associated confidence regions.

The purpose of this work is to gain additional insight into intuitive understanding of how the CBPKF solution differs from the KF, by casting them in error and state spaces using a bi-state model, and to advance understanding of the comparative performance of the CBEnKF with the EnKF by identifying and characterizing the representative geometric attributes in the state space via eigenvalue analysis. The new and significant contributions of this work are: comparative geometric analysis of the CBPKF solution with the KF, development of a set of geometry-based relationships for improved understanding of the CBPKF solution, and geometric characterization of ensemble analysis for improved understanding of the comparative performance of the CBEnKF with the EnKF. The comparative performance is based on numerical experiments with the Lorenz 63 model, chosen in this study for familiarity and simplicity. For real-world flood forecasting applications of the CBEnKF, the interested reader is referred to Lee et al. [5] and Shen et al. [18].

This paper is organized as follows. Section 2 describes the state updating problem, and the two solution approaches, the KF and the CBPKF. Section 3 presents the geometric analysis of the CB-informed KF in relation to the KF and the EnKF. First, Section 3.1 describes how to use vectors to represent stochastic variables in the Euclidean domain, and Section 3.2 expresses the filter equations in terms of state and observation error terms. Section 3.3 geometrically illustrates the filter equations for a low-order model based on the visualization of error vectors and their relations to Kalman gain and covariance matrices in the Euclidean space [19]. Section 3.4 describes geometric analyses in the state-space with eigenvalues and eigenvectors of error covariance matrices and associated confidence regions. In Section 3.5, an example is given for geometric analysis in the state space in which the CBEnKF, the EnKF, and Open Loop (OL) are compared for the Lorenz 63 model via numerical experiments. Finally, Section 4 describes conclusions.

## 2. Methodology

### 2.1. State Updating Problem

The nonlinear dynamical model is written as Equation (1).

$$X_k = M(X_{k-1}) + W_{k-1} \tag{1}$$

In the above, $X_k$ denotes the $(n_c \times 1)$ model state, or control vector, where $n_c$ denotes the number of variables in the control vector, M( ) denotes the dynamical model for state variables, and $W_{k-1}$ denotes the dynamical model error at a time step $k-1$ with a mean of zero and a covariance of $Q_k$. The nonlinear observational model is written as:

$$Z_k = H_k(X_k) + V_k \tag{2}$$

In the above, $Z_k$ denotes the (n × 1) observation vector, where n denotes the total number of observations. $V_k$ denotes the (n × 1) observation error vector at a time step

$k$ with a mean of zero and a covariance of $R_k$. $H_k()$ denotes the nonlinear observation operator that maps $X_k$ to $Z_k$. This study solves the linear observation model which renders $H_k(X_k)$ in Equation (2) linear, i.e., $H_k X_k$:

$$Z_k = H_k X_k + V_k \tag{3}$$

In the case of a nonlinear observation operator, e.g., soil moisture–streamflow transformation, one can still render Equation (2) linear via state augmentation [5,20,21], which is beyond the scope of this study. Interested readers are referred to Lee et al. [5], where state augmentation is used to solve a highly nonlinear flood forecasting problem.

Equation (4) shows the state updating equation where $X_{k|k}$, $X_{k|k-1}$, and $K_k$ represent updated state at a time step $k$, state forecast from $k-1$ to $k$, and Kalman gain, respectively.

$$X_{k|k} = X_{k|k-1} + K_k \left[ Z_k - H_k X_{k|k-1} \right] \tag{4}$$

To find $K_k$, the KF and the EnKF minimize the error variance ($\Sigma_{EV}$) in Equation (5). The CBPKF and the CBEnKF minimize the weighted sum of $\Sigma_{EV}$ and the expectation of the Type-II CB squared ($\Sigma_{CB}$) in Equation (6), or $\Sigma_{EV} + \alpha \Sigma_{CB}$ where $\alpha$ is the weight given to the *CB* penalty term. The $\alpha$ can be estimated using an iterative method that yields the error covariance within theoretically expected bounds (see [5] for details). Since the iterative method is computationally expensive, the adaptive filtering method has been developed, and the results are reported in Shen et al. [17].

$$\Sigma_{EV} = E \left[ \left( X_k - X_{k|k} \right) \left( X_k - X_{k|k} \right)^T \right] \tag{5}$$

$$\Sigma_{CB} = E[(X_k - E_{X_{k|k}}[X_{k|k}|X_k])(X_k - E_{X_{k|k}}[X_{k|k}|X_k])^T] \tag{6}$$

The following sections present expressions for Kalman gain ($K_k$) and covariance ($P_{k|k}$) matrices in the case of the KF and the CBPKF.

### 2.2. Kalman Filter, KF

Equation (4) is rewritten as Equation (7) but with the superscript $K$ to denote the KF.

$$X_{k|k}^K = X_{k|k-1} + K_k^K \left[ Z_k - H_k X_{k|k-1} \right] \tag{7}$$

Minimizing $\Sigma_{EV}$ in Equation (5) results in Kalman gain $K_k^K$ in Equation (8) and the covariance analysis matrix $P_{k|k}$ in Equation (9) [1].

$$K_k^K = P_{k|k-1} H^T \left[ H P_{k|k-1} H^T + R \right]^{-1} \tag{8}$$

$$P_{k|k} = [1 - K_k H] P_{k|k-1} = P_{k|k-1} - K_k \left[ H P_{k|k-1} H^T + R_k \right] K_k^T \tag{9}$$

### 2.3. Conditional Bias-Penalized Kalman Filter, CBPKF

State updating equation (Equation (4)) can be rewritten into Equation (10) with the superscript $C$ denoting the CBPKF.

$$X_{k|k}^C = X_{k|k-1} + K_k^C \left[ Z_k - H_k X_{k|k-1} \right] \tag{10}$$

Minimizing $\Sigma_{EV} + \alpha \Sigma_{CB}$ produces the CBPKF gain $K_k^C$ in Equation (11) and the covariance matrix $\Sigma_{k|k}$ in Equation (12) [3,4,17].

$$K_k^C = [\varpi_{1,k} H_k + \varpi_{2,k}]^{-1} \varpi_{1,k} \tag{11}$$

$$\Sigma_{k|k} = [\omega_{1,k} H_k + \omega_{2,k}]^{-1} \left( \omega_{1,k} R_k \omega_{1,k}^T + \omega_{2,k} \Sigma_{k|k-1} \omega_{2,k}^T \right) [\omega_{1,k} H_k + \omega_{2,k}]^{-1} \qquad (12)$$

In Equations (11) and (12), the (m × n) and (m × m) weight matrices for the observations and model prediction, $\omega_{1,k}$ and $\omega_{2,k}$, respectively, are given by:

$$\omega_{1,k} = \hat{H}_k^T \Gamma_{11,k} + \Gamma_{21,k} \qquad (13)$$

$$\omega_{2,k} = \hat{H}_k^T \Gamma_{12,k} + \Gamma_{22,k} \qquad (14)$$

In Equations (13) and (14), the (m × n) modified observation matrix, $\hat{H}_k^T$, and the (n × n), (n × m), and (m × m) matrices, $\Gamma_{11,k}$, $\Gamma_{12,k}$ $\left( = \Gamma_{21,k}^T \right)$, and $\Gamma_{22,k}$, respectively, are given by:

$$\hat{H}_k^T = H_k^T + \alpha C_k^T \qquad (15)$$

$$\Gamma_{22,k}^{-1} = \Lambda_{22,k} - \Lambda_{21,k} \Lambda_{11,k}^{-1} \Lambda_{12,k} \qquad (16)$$

$$\Gamma_{11,k} = \Lambda_{11,k}^{-1} + \Lambda_{11,k}^{-1} \Lambda_{12,k} \Gamma_{22,k} \Lambda_{21,k} \Lambda_{11,k}^{-1} \qquad (17)$$

$$\Gamma_{12,k} = -\Lambda_{11,k}^{-1} \Lambda_{12,k} \Gamma_{22,k} \qquad (18)$$

In Equations (15)–(18), the (n × m) CB gain matrix for the observation vector, $C_k$, and the (n × n), (n × m), and (m × m) modified error covariance matrices, $\Lambda_{11,k}$, $\Lambda_{12,k}$ $\left( = \Lambda_{12,k}^T \right)$, and $\Lambda_{22,k}$, respectively, are given by:

$$C_k = \left[ H_k \Sigma_{k|k-1} G_{2,k}^{-1} + R_k H_k \right] \left[ G_{2,k}^{-1} \Sigma_{k|k-1} G_{2,k}^{-1} + 2(H_k^T R_k H_k + \Sigma_{k|k-1}) \right]^{-1} G_{2,k}^{-1} \qquad (19)$$

$$\Lambda_{11,k} = R_k + \alpha(1-\alpha) C_k \Sigma_{k|k-1} C_k^T + \Lambda_{12,k} H_k^T + H_k \Lambda_{21,k} \qquad (20)$$

$$\Lambda_{12,k} = -\alpha C_k \Sigma_{k|k-1} \qquad (21)$$

$$\Lambda_{22,k} = \Sigma_{k|k-1} \qquad (22)$$

In Equations (19)–(22), $R_k$ and $\Sigma_{k|k-1}$ denote the (n × n) observation error covariance matrix, $Cov[V_k, V_k]$, and the (m × m) forecast error covariance matrix, respectively, and $G_{2,k}^{-1} = H_k^T H_k + I$.

If $\alpha = 0$, the CBPKF solution is reduced to the KF, i.e., $\Sigma_{k|k} = \left[ H_k^T R_k^{-1} H_k + \Sigma_{k|k-1}^{-1} \right]^{-1}$, and $K_k = \Sigma_{k|k} H_k^T R_k^{-1}$ [5].

## 3. Geometric Analysis of the CB-Informed KF

### 3.1. Error Representation of Filter Equations

Since the error covariance of stochastic variables plays the key role in geometric analyses in the Euclidean space, this section defines errors in states and a residual in measurement in order to rewrite state updating and observation equations with error terms.

$$\varepsilon_k^C = X_k - X_{k|k}^C \qquad (23)$$

$$\varepsilon_k^K = X_k - X_{k|k}^K \qquad (24)$$

$$\varepsilon_k = X_k - X_{k|k-1} \qquad (25)$$

$$y_k = Z_k - H_k X_{k|k-1} \qquad (26)$$

In the above, $\varepsilon_k^C$, $\varepsilon_k^K$, and $\varepsilon_k$ denote the CBPKF analysis error, the KF analysis error, and the forecast error, respectively; $y_k$ represents innovation; $X_k$, $X_{k|k}^C$, $X_{k|k}^K$, and $X_{k|k-1}$ represent the truth, the CBPKF-updated state, the KF-updated state, and the state forecast, respectively.

State update equations for the KF and the CBPKF are rewritten into Equations (27) and (28), assuming the same a priori states used in both filters.

$$\varepsilon_k^K = \varepsilon_k - K_k^K y_k \tag{27}$$

$$\varepsilon_k^C = \varepsilon_k - K_k^C y_k \tag{28}$$

where $K_k^K$ and $K_k^C$ are Kalman gains from the KF and the CBPKF, respectively.

Updated states from both filters have the following mathematical relationship.

$$\varepsilon_k^C = \varepsilon_k^K - \left(K_k^C - K_k^K\right) y_k \tag{29}$$

The observation model in Equation (3) can be rewritten as follows:

$$y_k = H_k \varepsilon_k + V_k \tag{30}$$

From Equation (3), the covariance matrix of $y_k$, or $D_k$, can be written as

$$D_k = H_k P_k^- H_k^T + R_k \tag{31}$$

where $P_k^-$ is the covariance matrix of the state forecast.

*3.2. KF and CBPKF Solutions for a Bi-State Model*

A bi-state model ($X_k = [x_{1,k} \; x_{2,k}]^T$) is used to illustrate the geometric relation of errors in states and observations from the CBPKF and the KF. Observation $z_{1,k}$ exists for a state $x_{1,k}$ but not for $x_{2,k}$ to investigate how state updating can be illustrated in the two-dimensional (2-D) Euclidean space for observable as well as unobservable states, i.e., $Z_k = [z_{1,k}]$; $H = [1 \; 0]$ is used for simplicity. Kronhamn [19] used the same $X_k$, $Z_k$ and $H$ matrices as those described in this Section for the geometric illustration of the KF, whereas we focus on the CBPKF in reference to the KF. For the bi-state model with a forecast error covariance $P_{f,k} = \begin{bmatrix} \sigma_{1,k}^2 & \sigma_{12,k} \\ \sigma_{12,k} & \sigma_{2,k}^2 \end{bmatrix}$ and the observation error $R_k = \left[\sigma_{z,k}^2\right]$, matrices of Kalman gain (Equations (32) and (33)), analysis error covariance (Equations (34) and (35)), and correlation (Equations (36)–(38)) for the CBPKF and the KF are evaluated in the equations below where $\sigma_{12,k} = \rho_{12,k}\sigma_{1,k}\sigma_{2,k}$ with $\rho_{12,k}$ representing the correlation between $x_{1,k}$ and $x_{2,k}$.

$$K_k^C = \begin{bmatrix} K_{1,k}^C \\ K_{2,k}^C \end{bmatrix} = K_k^K + \alpha \begin{bmatrix} \dfrac{2\sigma_{1,k}^2 \sigma_{z,k}^2}{\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right)\left((1+2\alpha)\sigma_{1,k}^2 + \sigma_{z,k}^2\right)} \\[3mm] \dfrac{2\sigma_{12,k}\sigma_{z,k}^2}{\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right)\left((1+2\alpha)\sigma_{1,k}^2 + \sigma_{z,k}^2\right)} \end{bmatrix} \tag{32}$$

where

$$K_k^K = \begin{bmatrix} K_{1,k}^K \\ K_{2,k}^K \end{bmatrix} = \begin{bmatrix} \dfrac{\sigma_{1,k}^2}{\sigma_{1,k}^2 + \sigma_{z,k}^2} \\[3mm] \dfrac{\sigma_{12,k}}{\sigma_{1,k}^2 + \sigma_{z,k}^2} \end{bmatrix} \tag{33}$$

Equation (32) indicates that for an observable state $x_{1,k}$, $K_k^C$ is always larger than $K_k^K$, but for an unobservable state $x_{2,k}$, the sign of $K_k^C - K_k^K$ depends on the sign of $\sigma_{12,k}$. The CBPKF covariance matrix $P_{a,k}^C$ and its relation to the KF-equivalent $P_{a,k}^K$ are given in Equation (34).

$$P_{a,k}^C = P_{a,k}^K + \alpha \begin{bmatrix} \dfrac{\alpha\left(2\sigma_{1,k}^2\sigma_{z,k}^2\right)^2}{\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right)\left((1+2\alpha)\sigma_{1,k}^2 + \sigma_{z,k}^2\right)^2} & \dfrac{\alpha\sigma_{12,k}\left(2\sigma_{1,k}^2\sigma_{z,k}^2\right)^2}{\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right)\left((1+2\alpha)\sigma_{1,k}^2 + \sigma_{z,k}^2\right)^2} \\[4mm] \dfrac{\alpha\sigma_{12,k}\left(2\sigma_{1,k}^2\sigma_{z,k}^2\right)^2}{\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right)\left((1+2\alpha)\sigma_{1,k}^2 + \sigma_{z,k}^2\right)^2} & \dfrac{\alpha\left(2\sigma_{12,k}\sigma_{z,k}^2\right)^2}{\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right)\left((1+2\alpha)\sigma_{1,k}^2 + \sigma_{z,k}^2\right)^2} \end{bmatrix} \tag{34}$$

where

$$P_{a,k}^K = \begin{bmatrix} \frac{\sigma_{1,k}^2 \sigma_{z,k}^2}{\sigma_{1,k}^2 + \sigma_{z,k}^2} & \frac{\sigma_{12,k} \sigma_{z,k}^2}{\sigma_{1,k}^2 + \sigma_{z,k}^2} \\ \frac{\sigma_{21,k} \sigma_{z,k}^2}{\sigma_{1,k}^2 + \sigma_{z,k}^2} & \sigma_{2,k}^2 - \frac{\sigma_{12,k}^2}{\sigma_{1,k}^2 + \sigma_{z,k}^2} \end{bmatrix} \tag{35}$$

In Equation (34), variances of both updated states $x_{1,k}$ and $x_{2,k}$ are larger than KF-equivalents, and the sign of covariance terms of $P_{a,k}^C - P_{a,k}^K$ depend on the sign of $\sigma_{12,k}$. Equations (34) and (35) imply that the CBPKF-updated state ensembles have larger spreads in the state space than the KF-updated, and that the matrix norm of $P_{a,k}^C$ is larger than that of $P_{a,k}^K$. From covariance matrices, Pearson product-moment correlation matrix $C_{a,k}^C$ can be computed by Equation (36).

$$C_{a,k}^C = \left( diag\left(P_{a,k}^C\right) \right)^{-\frac{1}{2}} P_{a,k}^C \left( diag\left(P_{a,k}^C\right) \right)^{-\frac{1}{2}} = \begin{bmatrix} 1 & C_{a,12,k}^C \\ C_{a,12,k}^C & 1 \end{bmatrix} \tag{36}$$

where *diag*( ) denotes a diagonal matrix. Equation (37) shows that the correlation coefficient of the CBPKF-updated states $C_{a,12,k}^C$ can be either larger or smaller than the KF-equivalent $C_{a,12,k}^K$.

$$
\begin{aligned}
C_{a,12,k}^C &= \frac{\sigma_{12,k}\sigma_{z,k}\sqrt{\left((1+2\alpha)^2\sigma_{1,k}^2 + \sigma_{z,k}^2\right)}}{\sigma_{1,k}\sqrt{\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right)\left(\sigma_{2,k}^2\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right) - \sigma_{12,k}^2\right) + 4\alpha(1+\alpha)\left(\sigma_{1,k}^2 - \sigma_{z,k}^2\right)\left(\sigma_{1,k}^2\sigma_{2,k}^2 - \sigma_{12,k}^2\right)}} \\
&= C_{a,12,k}^K \\
&\quad + \frac{\sigma_{12,k}\sigma_{z,k}\sqrt{A_1 - A_2}}{\sigma_{1,k}\sqrt{\sigma_{2,k}^2\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right) - \sigma_{12,k}^2}\sqrt{\left(\sigma_{2,k}^2\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right) - \sigma_{12,k}^2\right)\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right) + 4\alpha(1+\alpha)\left(\sigma_{1,k}^2 - \sigma_{z,k}^2\right)\left(\sigma_{1,k}^2\sigma_{2,k}^2 - \sigma_{12,k}^2\right)}}
\end{aligned} \tag{37}
$$

where

$$A_1 = \left(\sigma_{2,k}^2\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right) - \sigma_{12,k}^2\right)\left((1+2\alpha)^2\sigma_{1,k}^2 + \sigma_{z,k}^2\right)$$

$$A_2 = \left(\sigma_{2,k}^2\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right) - \sigma_{12,k}^2\right)\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right) + 4\alpha(1+\alpha)\left(\sigma_{1,k}^2 - \sigma_{z,k}^2\right)\left(\sigma_{1,k}^2\sigma_{2,k}^2 - \sigma_{12,k}^2\right)$$

$$C_{a,12,k}^K = \frac{\sigma_{12,k}\sigma_{z,k}}{\sigma_{1,k}\sqrt{\sigma_{2,k}^2\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right) - \sigma_{12,k}^2}} \tag{38}$$

From Equations (32)–(38), if $\alpha = 0$, then $K_k^C = K_k^K$, $P_{a,k}^C = P_{a,k}^K$, and $C_{a,k}^C = C_{a,k}^K$.

### 3.3. Geometric Representation of KF and CBPKF Solutions

This Section begins with describing a relation between stochastic variables and vectors in the Euclidean domain at Equations (39)–(41), and then describes the geometric representation of KF and CBPKF equations. The covariance of stochastic variables $x_s$ and $y_s$ can be used to compute the scalar product of two vectors $\vec{x_E}$ and $\vec{y_E}$ in the Euclidean domain (Equation (39)). The vector norm $||\vec{x_E}||$ corresponds to the standard deviation of $x_s$ (Equation (40)). The angle of $\vec{x_E}$ and $\vec{y_E}$ can be computed from the correlation of $x_s$ and $y_s$ (Equation (41)) [19,22]. Figure 1 shows the vector representation of stochastic variables in the Euclidean space.

$$\vec{x_E} \cdot \vec{y_E} = ||\vec{x_E}|| \, ||\vec{y_E}|| \cos\theta = Cov[x_s, y_s] = \sigma_{x_s y_s} \tag{39}$$

$$||\vec{x_E}|| = \sqrt{\vec{x_E} \cdot \vec{x_E}} = \sqrt{Cov[x_s, x_s]} = \sigma_{x_s} \tag{40}$$

$$\theta = \cos^{-1}\left(\frac{\vec{x_E} \cdot \vec{y_E}}{||\vec{x_E}|| \, ||\vec{y_E}||}\right) = \cos^{-1}\left(\frac{\sigma_{x_s y_s}}{\sigma_{x_s}\sigma_{y_s}}\right) = \cos^{-1}\left(\rho_{x_s y_s}\right) \tag{41}$$
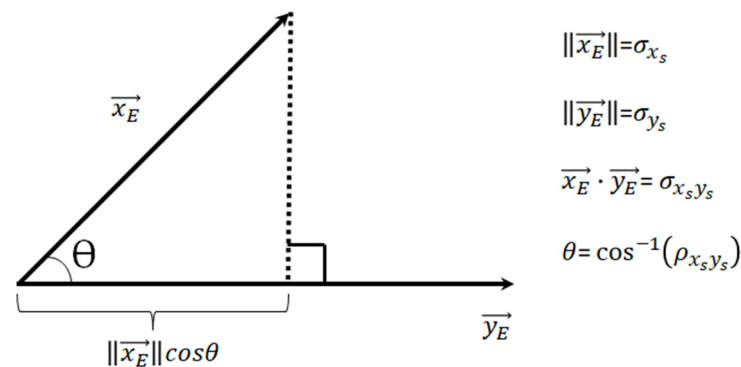
**Figure 1.** Vector representation of stochastic variables in the Euclidean space, where $x_E$ and $y_E$ are vectors in the Euclidean space and $x_S$ and $y_S$ are stochastic variables.

Figure 2 shows the error vectors from (a) the KF and (b) the CBPKF for the observable state $x_{1,k}$ of the bi-state model, where $||\vec{\varepsilon}_{1,k}|| = \sigma_{1,k}$, $||\vec{v}_{1,k}|| = \sigma_{z,k}$, and $||\vec{y}_{1,k}||^2 = \left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right)$, where the last equality is from Equation (30). In Figure 2a, the state forecast error vector $\vec{\varepsilon}_{1,k}$ is orthogonal to the observation error vector $\vec{v}_{1,k}$ owing to the independence assumption. Being a minimum variance solution, $\vec{\varepsilon}_{1,k}^{K}$ is orthogonal to $\vec{y}_{1,k}^{K}$ [19,23]. Figure 2a also shows that the forecast error is a vector sum of the gain-weighted innovation and the analysis error, i.e., $\vec{\varepsilon}_{1,k} = K_{1,k}^{K}\vec{y}_{1,k} + \vec{\varepsilon}_{1,k}^{K}$ as expected from Equation (27). The KF analysis error variance may be obtained in Figure 2a via the Pythagorean theorem as $||\vec{\varepsilon}_{1,k}^{K}||^2 = ||\vec{\varepsilon}_{1,k}||^2 - ||K_{1,k}^{K}\vec{y}_{1,k}||^2 = \sigma_{1,k}^2 - \left(\frac{\sigma_{1,k}^2}{\sigma_{1,k}^2+\sigma_{z,k}^2}\right)^2\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right) = \frac{\sigma_{1,k}^2\sigma_{z,k}^2}{\sigma_{1,k}^2+\sigma_{z,k}^2}$ where $K_{1,k}^{K} = \frac{\sigma_{1,k}^2}{\sigma_{1,k}^2+\sigma_{z,k}^2}$ as expected from Equation (35). Since $\vec{\varepsilon}_{1,k}^{C} = \vec{\varepsilon}_{1,k}^{K} - \left(K_{1,k}^{C} - K_{1,k}^{K}\right)\vec{y}_{1,k}$, we may write via the Pythagorean theorem $||\vec{\varepsilon}_{1,k}^{C}|| = \sqrt{||\vec{\varepsilon}_{1,k}^{K}||^2 + ||\left(K_{1,k}^{C} - K_{1,k}^{K}\right)\vec{y}_{1,k}||^2} = \sqrt{\frac{\left((1+2\alpha)^2\sigma_{1,k}^2+\sigma_{z,k}^2\right)\sigma_{1,k}^2\sigma_{z,k}^2}{\left((1+2\alpha)\sigma_{1,k}^2+\sigma_{z,k}^2\right)^2}}$ which is expected from Equation (34). In Figure 2, the inequality, $||K_{1,k}^{K}\vec{y}_{1,k}|| < ||K_{1,k}^{C}\vec{y}_{1,k}||$, arises due to the fact that the CBPKF solution minimizes not the error variance but a weighted sum of the error variance and the variance of the Type-II CB.

Figure 3 is the same as Figure 2 but for the unobservable state $x_{2,k}$ where $||\vec{\varepsilon}_{2,k}|| = \sigma_{2,k}$; $||\vec{\varepsilon}_{2p,k}|| = ||\vec{\varepsilon}_{2,k}||\cos\theta = \frac{\sigma_{12,k}}{\sigma_{1,k}}$; $||\vec{\varepsilon}_{2o,k}|| = \sqrt{||\vec{\varepsilon}_{2,k}||^2 - ||\vec{\varepsilon}_{2p,k}||^2} = \sqrt{\sigma_{2,k}^2 - \frac{\sigma_{12,k}^2}{\sigma_{1,k}^2}}$; and $||\vec{\varepsilon}_{2c,k}|| = \sqrt{||\vec{\varepsilon}_{2p,k}||^2 - ||K_{2,k}^{K}\vec{y}_{1,k}||^2} = \sqrt{\frac{\sigma_{12,k}^2\sigma_{z,k}^2}{\sigma_{1,k}^2\left(\sigma_{1,k}^2+\sigma_{z,k}^2\right)}}$. In Figure 3a, the proportionality, $||\vec{\varepsilon}_{2p,k}|| : ||K_{2,k}^{K}\vec{y}_{1,k}|| = ||\vec{\varepsilon}_{1,k}|| : ||K_{1,k}^{K}\vec{y}_{1,k}||$, gives $K_{2,k}^{K} = K_{1,k}^{K}\frac{||\vec{\varepsilon}_{2p,k}||}{||\vec{\varepsilon}_{1,k}||} = K_{1,k}^{K}\frac{||\vec{\varepsilon}_{2,k}||\cos\theta}{||\vec{\varepsilon}_{1,k}||} = K_{1,k}^{K}\frac{\sigma_{2,k}}{\sigma_{1,k}}\frac{\sigma_{12,k}}{\sigma_{1,k}\sigma_{2,k}} = K_{1,k}^{K}\frac{\sigma_{12,k}}{\sigma_{1,k}^2}$ as expected from Equation (33). In Figure 3a, the orthogonality, $\vec{\varepsilon}_{2,k}^{K} \perp \vec{y}_{1,k}^{K}$, and the equality $\vec{\varepsilon}_{2,k}^{K} = \vec{\varepsilon}_{2,k} - K_{2,k}^{K}\vec{y}_{1,k}$, yield $||\vec{\varepsilon}_{2,k}^{K}||^2 = ||\vec{\varepsilon}_{2,k}||^2 - ||K_{2,k}^{K}\vec{y}_{1,k}||^2 = \sigma_{2,k}^2 - \frac{\sigma_{12,k}^2}{\sigma_{1,k}^2+\sigma_{z,k}^2}$ in agreement with Equation (35). In Figure 3b, $||\vec{\varepsilon}_{2,k}^{C}||^2$ may be written via the Pythagorean theorem as:

$$||\vec{\varepsilon}_{2,k}^{C}||^2 = ||\vec{\varepsilon}_{2,k}^{K}||^2 + ||\left(K_{2,k}^{C} - K_{2,k}^{K}\right)\vec{y}_{1,k}||^2 \tag{42}$$
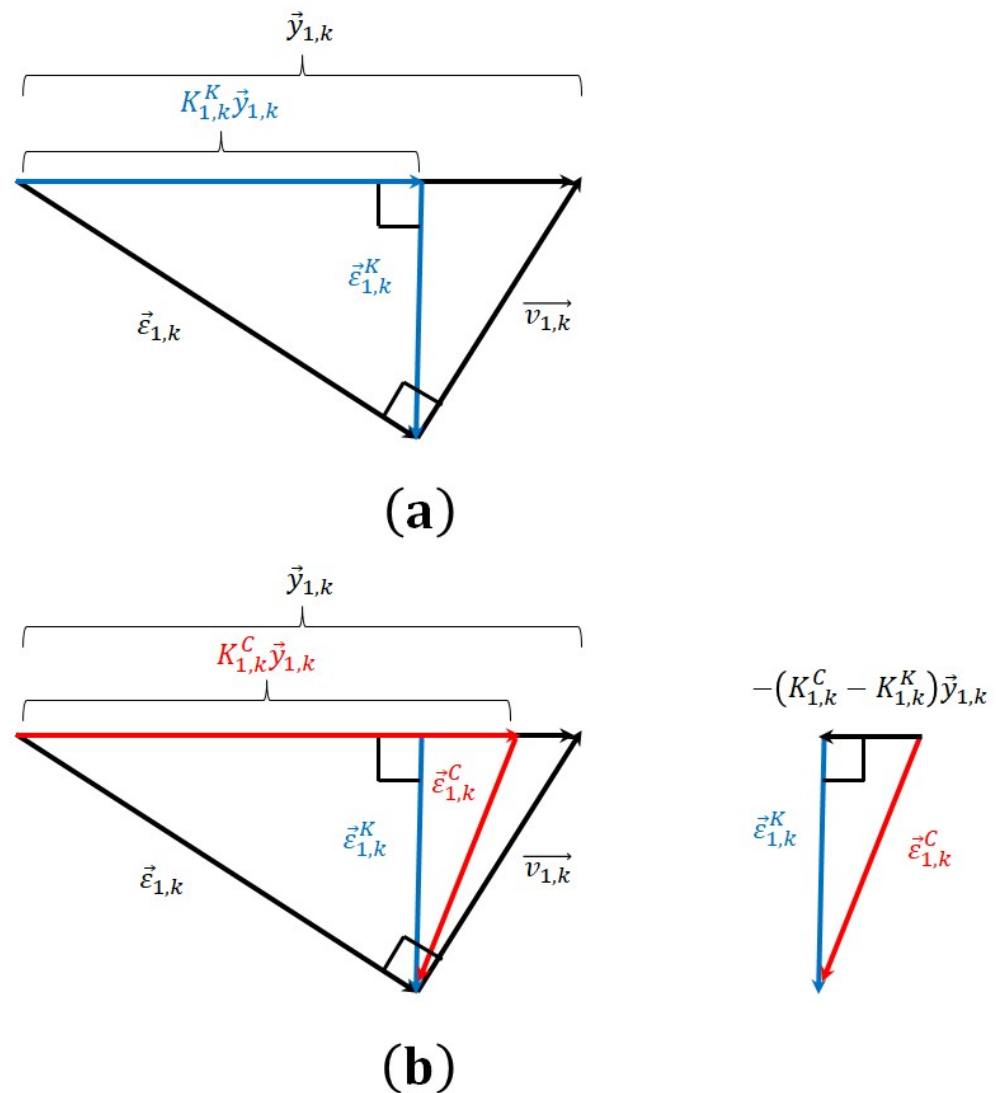
**Figure 2.** Geometric illustration of assimilation solutions for the observable state $x_{1,k}$ for (**a**) the KF ($\rightarrow$; after Kronhamn [19]), and (**b**) the CBPKF ($\rightarrow$).

Using Equations (35), (32) and (31) for $\left|\left|\vec{\varepsilon}_{2,k}^{K}\right|\right|^{2}$, $K_{2,k}^{C} - K_{2,k}^{K}$, and $\left|\left|\vec{y}_{1,k}\right|\right|^{2}$, respectively, we may rewrite Equation (42), in agreement with Equation (34), as:

$$\left|\left|\vec{\varepsilon}_{2,k}^{C}\right|\right|^{2} = \sigma_{2,k}^{2} - \frac{\sigma_{12,k}^{2}}{\sigma_{1,k}^{2} + \sigma_{z,k}^{2}} + \frac{\left(2\alpha\sigma_{12,k}\sigma_{z,k}^{2}\right)^{2}}{\left(\sigma_{1,k}^{2} + \sigma_{z,k}^{2}\right)\left((1 + 2\alpha)\sigma_{1,k}^{2} + \sigma_{z,k}^{2}\right)^{2}} \tag{43}$$

Figures 2 and 3 show that the KF- and the CBPKF-updated state error vectors point to different directions in the state space. Figure 4 shows the updated state error vectors in Figures 2 and 3 to visually compare the differences in the angle, the magnitude, and the direction. The angles of the two-state error vectors for the CBPKF ($\theta^{C}$ in Equation (44)) and the KF ($\theta^{K}$ in Equation (45)) can be computed from the correlation $C_{a,12,k}^{C}$ and $C_{a,12,k}^{K}$ in Equations (37) and (38), respectively:

$$\theta^{C} = \cos^{-1} C_{a,12,k}^{C} = \cos^{-1}\left(\frac{\vec{\varepsilon}_{1,k}^{C} \cdot \vec{\varepsilon}_{2,k}^{C}}{\left|\left|\vec{\varepsilon}_{1,k}^{C}\right|\right| \left|\left|\vec{\varepsilon}_{2,k}^{C}\right|\right|}\right) = \cos^{-1}\left(\frac{\sigma_{x_{1,k|k}^{C}} x_{2,k|k}^{C}}{\sigma_{x_{1,k|k}^{C}} \sigma_{x_{2,k|k}^{C}}}\right) \tag{44}$$

$$\theta^K = \cos^{-1} C_{a,12,k}^K = \cos^{-1}\left(\frac{\overrightarrow{\varepsilon}_{1,k}^K \cdot \overrightarrow{\varepsilon}_{2,k}^K}{||\overrightarrow{\varepsilon}_{1,k}^K||\,||\overrightarrow{\varepsilon}_{2,k}^K||}\right) = \cos^{-1}\left(\frac{\sigma_{x_{1,k}^K x_{2,k}^K}}{\sigma_{x_{1,k|k}^K} \sigma_{x_{2,k|k}^K}}\right) \tag{45}$$
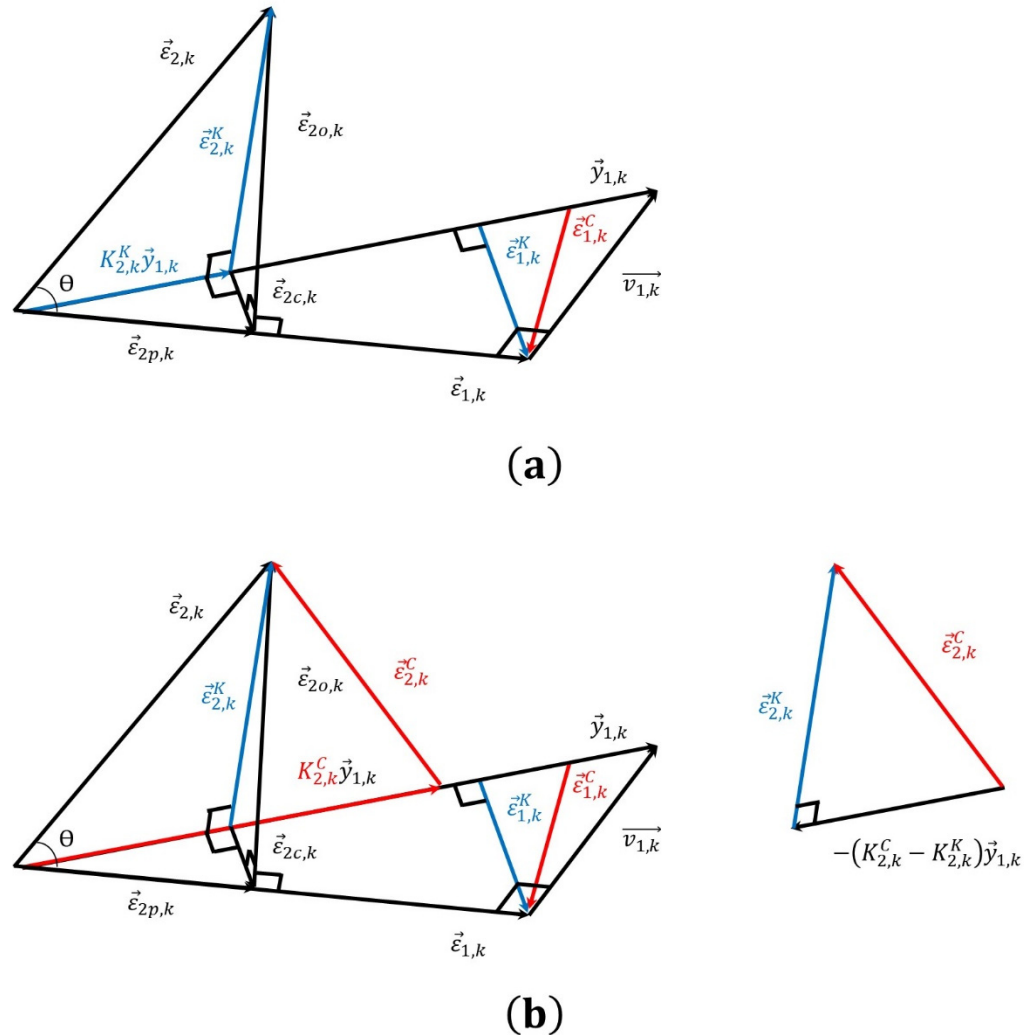


(**a**)



(**b**)

**Figure 3.** Geometric illustration of assimilation solutions for the unobservable state $x_{2,k}$ for (**a**) the KF (→; after Kronhamn [19]), and (**b**) the CBPKF (→).

Below, we develop a set of geometric expressions in the 2-D state space for the analysis error covariance via eigenvalue decomposition (EVD, [24]).

### 3.4. Geometric Analysis in the State Space

Geometric characteristics of state ensembles in the 2-D state space can be quantified by confidence regions (CR), eigenvectors, eigenvalues, and the angle between the eigenvector and the basis vector of the x-axis. Assuming normal distributions for state ensembles in a 2-D state space, a CR, or so-called a covariance ellipse, can be constructed based on the EVD of a covariance matrix as well as the Chi-Square probability table. The presence of the CB results in different variances (eigenvalues) and directions (eigenvectors) of updated state ensembles in a 2-D state space. The major and minor axis lengths of the covariance ellipse are $2\sqrt{s\lambda_1}$ and $2\sqrt{s\lambda_2}$ where $\lambda_1 > \lambda_2$ and the value of $s$ is from the Chi-Square probability table for a given confidence region, e.g., $s = 4.605$ for a 90% confidence region given the Chi-Square probability $P(s < 4.605) = 0.9$ in the case of degrees of freedom of 2. In a 2-D state space, the error in the orientation of the covariance ellipse with respect to the

truth can be estimated by the angle, $\theta$, between the largest eigenvector $\vec{u}_1$ and the vector $\vec{a}$ connecting the truth and the ensemble mean, i.e., $\theta\left(\vec{u}_1, \vec{a}\right) = \cos^{-1}\left(\frac{\vec{u}_1 \cdot \vec{a}}{||\vec{u}_1|| \; ||\vec{a}||}\right)$.
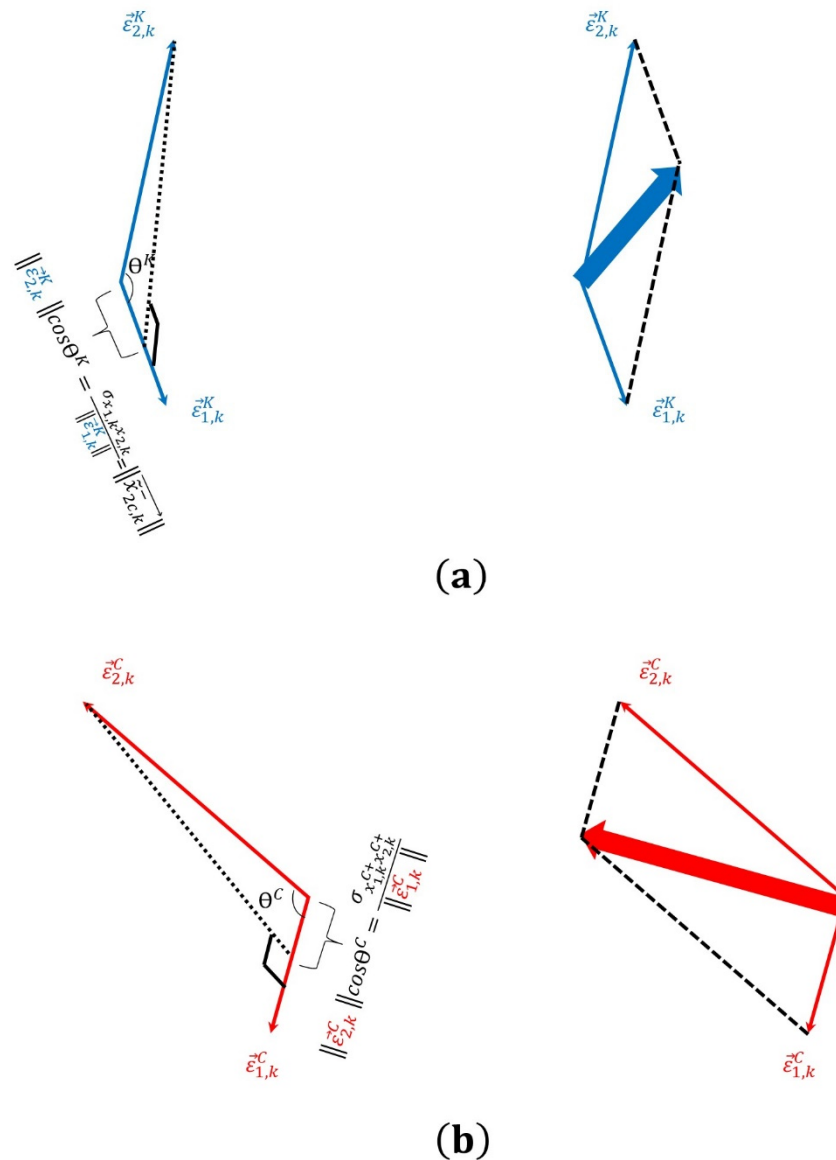


(**a**)



(**b**)

**Figure 4.** Vector representation of updated states from (**a**) the KF and (**b**) the CBPKF in Figure 3, where thick blue and red vectors indicate the principal error direction of the KF and the CBPKF in the Euclidean space, respectively.

The EVD of the CBPKF analysis covariance $P_{a,k}^C$ may be written as:

$$P_{a,k}^C = UEU^T \tag{46}$$

where

$$U = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} \vec{u}_1 & \vec{u}_2 \end{bmatrix}, \; E = \begin{bmatrix} \lambda_1^C & 0 \\ 0 & \lambda_2^C \end{bmatrix}, \tag{47}$$

In the above, $U$ is the eigenvector matrix which rotates the white data (W), or uncorrelated standard normal variates by $\theta$. The eigenvalue matrix E explains the variance along the principal error direction, or the direction of the eigenvector. In a 2-D state space, $\sqrt{E}$ is a scale factor applied to W. The dataset (D) resulting from scaling W by $\sqrt{E}$ and rotating by

$U$, i.e., $D = U\sqrt{E}W$, has the covariance matrix of $P_{a,k}^C = U\sqrt{E}\left(U\sqrt{E}\right)^T$. Below we apply EVD to the CBPKF analysis error covariance, $P_{a,k}^C$, from the bi-state model in Section 3.2.

With $\lambda_1^C > \lambda_2^C > 0$, and $P_{a,k}^K = \begin{bmatrix} p_{1,k}^2 & p_{12,k} \\ p_{12,k} & p_{2,k}^2 \end{bmatrix}$, $P_{a,k}^C$ in Equation (34) may be rewritten as:

$$P_{a,k}^C = P_{a,k}^K + \zeta \begin{bmatrix} \sigma_{1,k}^4 & \sigma_{12,k}\sigma_{1,k}^2 \\ \sigma_{12,k}\sigma_{1,k}^2 & \sigma_{12,k}^2 \end{bmatrix} \text{ where } \zeta = \frac{\left(2\alpha\sigma_{z,k}^2\right)^2}{\left(\sigma_{1,k}^2 + \sigma_{z,k}^2\right)\left((1+2\alpha)\sigma_{1,k}^2 + \sigma_{z,k}^2\right)^2} \quad (48)$$

Appendix A describes how eigenvalues and eigenvectors may be evaluated for a $2 \times 2$ covariance matrix. Using Equation (A11) in Appendix A, we have for $\lambda_1^C$:

$$\begin{aligned} \lambda_1^C &= 0.5\left(\left(p_{1,k}^2 + \zeta\sigma_{1,k}^4\right)^2 + \left(p_{2,k}^2 + \zeta\sigma_{12,k}^2\right)^2\right) \\ &+ 0.5\sqrt{\left(p_{1,k}^2 - p_{2,k}^2 + \zeta\sigma_{1,k}^4 - \zeta\sigma_{12,k}^2\right)^2 + 4\left(p_{12,k} + \zeta\sigma_{12,k}\sigma_{1,k}^2\right)^2} \end{aligned} \quad (49)$$

The difference in the largest eigenvalue between the KF and the CBPKF analysis error covariance is given by:

$$\begin{aligned} \lambda_1^C - \lambda_1^K &= 0.5\left(2p_{1,k}^2\zeta\sigma_{1,k}^4 + 2p_{2,k}^2\zeta\sigma_{12,k}^2 + \zeta^2\sigma_{1,k}^8 + \zeta^2\sigma_{12,k}^4\right) \\ &+ 0.5\sqrt{\left(p_{1,k}^2 - p_{2,k}^2 + \zeta\sigma_{1,k}^4 - \zeta\sigma_{12,k}^2\right)^2 + 4\left(p_{12,k} + \zeta\sigma_{12,k}\sigma_{1,k}^2\right)^2} \\ &- 0.5\sqrt{\left(p_{1,k}^2 - p_{2,k}^2\right)^2 + 4(p_{12,k})^2} \end{aligned} \quad (50)$$

where $\lambda_1^K$ denotes the largest eigenvalue of $P_{a,k}^K$. Using Equation (A12), we may write $\lambda_2^C$ and $\lambda_2^C - \lambda_2^K$:

$$\begin{aligned} \lambda_2^C &= 0.5\left(\left(p_{1,k}^2 + \zeta\sigma_{1,k}^4\right)^2 + \left(p_{2,k}^2 + \zeta\sigma_{12,k}^2\right)^2\right) \\ &- 0.5\sqrt{\left(p_{1,k}^2 - p_{2,k}^2 + \zeta\sigma_{1,k}^4 - \zeta\sigma_{12,k}^2\right)^2 + 4\left(p_{12,k} + \zeta\sigma_{12,k}\sigma_{1,k}^2\right)^2} \end{aligned} \quad (51)$$

$$\begin{aligned} \lambda_2^C - \lambda_2^K &= 0.5\left(2p_{1,k}^2\zeta\sigma_{1,k}^4 + 2p_{2,k}^2\zeta\sigma_{12,k}^2 + \zeta^2\sigma_{1,k}^8 + \zeta^2\sigma_{12,k}^4\right) \\ &- 0.5\sqrt{\left(p_{1,k}^2 - p_{2,k}^2 + \zeta\sigma_{1,k}^4 - \zeta\sigma_{12,k}^2\right)^2 + 4\left(p_{12,k} + \zeta\sigma_{12,k}\sigma_{1,k}^2\right)^2} \\ &+ 0.5\sqrt{\left(p_{1,k}^2 - p_{2,k}^2\right)^2 + 4(p_{12,k})^2} \end{aligned} \quad (52)$$

Similarly, using Equation (A14), we may write $\theta_{a,k}^C\left(\vec{u}_1, \vec{i}\right)$ and $\theta_{a,k}^C\left(\vec{u}_1, \vec{i}\right) - \theta_{a,k}^K\left(\vec{u}_1, \vec{i}\right)$ as:

$$\theta_{a,k}^C\left(\vec{u}_1, \vec{i}\right) = \frac{1}{2}\tan^{-1}\left(-\frac{2\left(p_{12,k} + \zeta\sigma_{12,k}\sigma_{1,k}^2\right)}{p_{1,k}^2 + \zeta\sigma_{1,k}^4 - p_{2,k}^2 - \zeta\sigma_{12,k}^2}\right) \quad (53)$$

$$\theta_{a,k}^C\left(\vec{u}_1, \vec{i}\right) - \theta_{a,k}^K\left(\vec{u}_1, \vec{i}\right) = \frac{1}{2}\tan^{-1}\left(-\frac{2\left(p_{12,k} + \zeta\sigma_{12,k}\sigma_{1,k}^2\right)}{p_{1,k}^2 + \zeta\sigma_{1,k}^4 - p_{2,k}^2 - \zeta\sigma_{12,k}^2}\right) - \frac{1}{2}\tan^{-1}\left(-\frac{2p_{12,k}}{p_{1,k}^2 - p_{2,k}^2}\right) \quad (54)$$

where the vector $\vec{i}$ is the basis vector of the x-axis and $\theta_{a,k}^C\left(\vec{u}_1, \vec{i}\right) - \theta_{a,k}^K\left(\vec{u}_1, \vec{i}\right)$ is equal to $\theta_{a,k}^C\left(\vec{u}_1, \vec{a}\right) - \theta_{a,k}^K\left(\vec{u}_1, \vec{a}\right)$. With the geometric attributes established above, we now carry out the comparative geometric analysis of the KF and the CBPKF analysis results using the Lorenz 63 model [25].

With the EVD of $P_{a,k}^C$ and the ensemble mean $(\overline{x_1}, \overline{x_2})$, the minimum percentage confidence $CR_{MIN}$ to contain the verifying truth $(x_{1,T}, x_{2,T})$ within the confidence region can be computed by $CR_{MIN} = 100 \times P(s < \widetilde{s})$, where $P(s < \widetilde{s})$ is the Chi-Square probability with degrees of freedom of 2; $\widetilde{s}$ satisfies Equation (55).

$$\frac{\left\{(x_{1,T} - \overline{x_1})\cos\theta_{a,k}^C + (x_{2,T} - \overline{x_2})\sin\theta_{a,k}^C\right\}^2}{\widetilde{s}\lambda_1^C} + \frac{\left\{(x_{1,T} - \overline{x_1})\sin\theta_{a,k}^C - (x_{2,T} - \overline{x_2})\cos\theta_{a,k}^C\right\}^2}{\widetilde{s}\lambda_2^C} = 1 \tag{55}$$

Figure 5 shows an example of computing $\vec{a}$, $\theta\left(\vec{u}_1, \vec{a}\right)$, $CR_{MIN}$, and eigenvalues and eigenvectors of a covariance matrix.
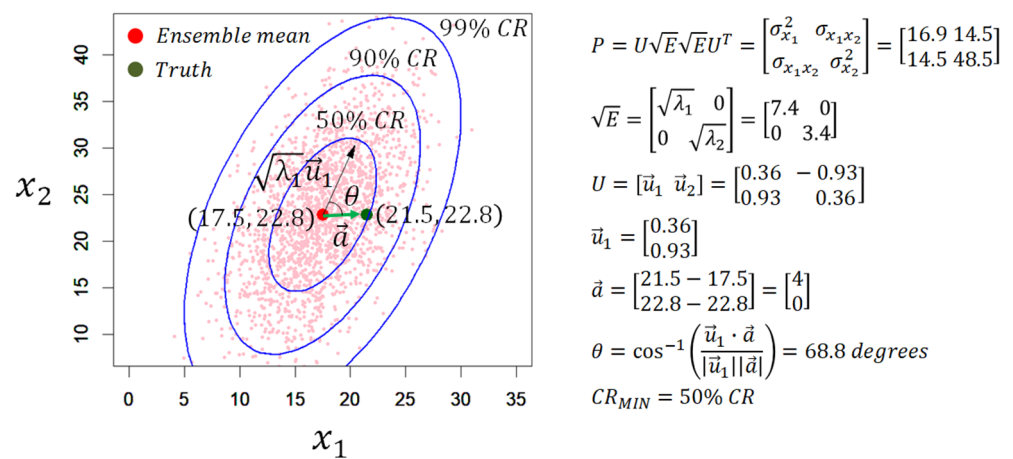


**Figure 5.** Illustration of geometric characteristics in a 2-D state space. Pink dots represent 2000 ensemble pairs of states $(x_1, x_2)$. The red and green dots denote the ensemble mean pair and the truth, respectively. $\vec{a}$ is the vector connecting the ensemble mean and the truth. $\vec{u}_1$ is the eigenvector of the covariance matrix $P$ that corresponds to the largest eigenvalue $\lambda_1$ of $P$. $\sqrt{\lambda_1}\vec{u}_1$ represents the vector along the major axis of covariance ellipses, where $\sqrt{\lambda_1}$ represents a scale factor applied to the white data. $\theta$ is the angle between $\vec{a}$ and $\vec{u}_1$. Assuming normal distributions of state ensembles, the major axis length of the covariance ellipse, or the confidence region (CR), is $2\sqrt{s\lambda_1}$, where $s$ = 1.39, 4.605 or 9.21 for 50, 90, or 99% CRs, respectively.

## 3.5. Numerical Experiment with the Lorenz 63 Model

In the sections above, a linear two-state model was used for theoretical simplicity. In this section, we use the three-state Lorenz 63 model to illustrate the differences between the EnKF and the CBEnKF solutions in terms of the geometric attributes introduced above. In this experiment, synthetically generated observations of all three states in the Lorenz 63 model were assimilated at every time step using the EnKF and the CBEnKF [26]. Preliminary experiments suggested observation error variances ($\sigma_z^2$) of 10 or 400 can be used for the cases of assimilating less uncertain or largely uncertain observations, respectively, based on the ensemble spread. To render the assimilation problem more challenging, $\sigma_z^2 = 400$ is used to compare the performance of the CBEnKF to that of the EnKF in Figures 6–11, where the ensemble size ($n_S$) used is 2000 to minimize filter performance degradation owing to a small ensemble size. Figures 12 and 13 compare assimilation results from the two filters in the case of $\sigma_z^2 = 10$ or 400 as a function of an ensemble size, where $n_S$ values used include 10, 20, 30, 50, 70, 100, 200, 300, 500, 700, 1000, and 2000.
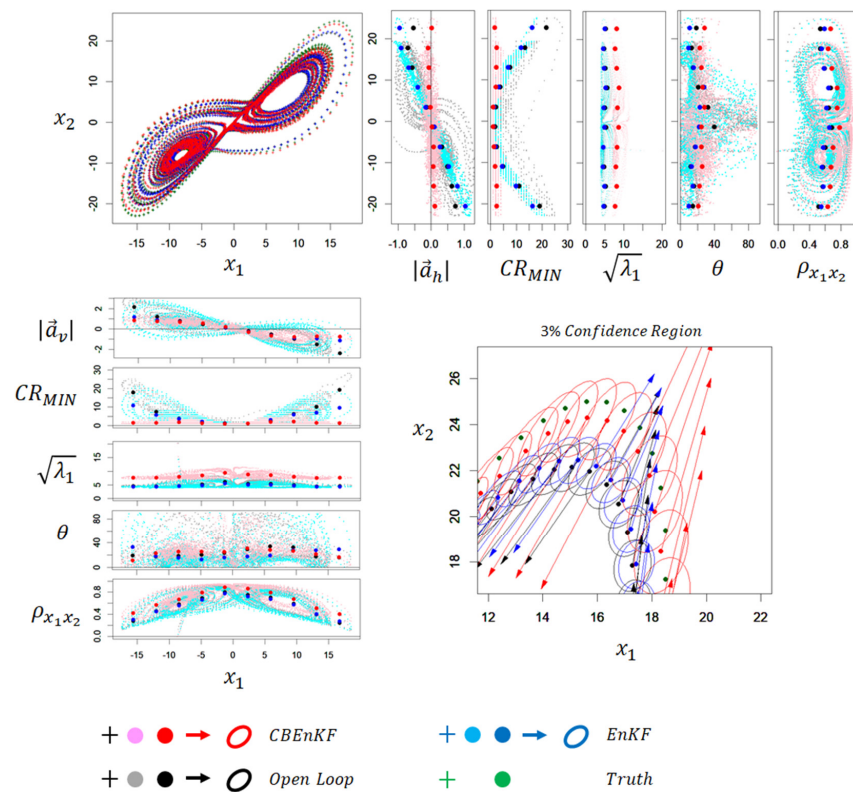
**Figure 6.** State space plot of $x_1$ and $x_2$ of the Lorenz 63 model from the CBEnKF, the EnKF, the Open Loop and the truth. Note that at the bottom right plot, $\vec{u}_1$ is plotted based on EVD results; however, $-\vec{u}_1$ (not shown) should also be considered to interpret the principal error vector pointing towards the truth—this is reflected in the shape and the size of covariance ellipses at the plot.
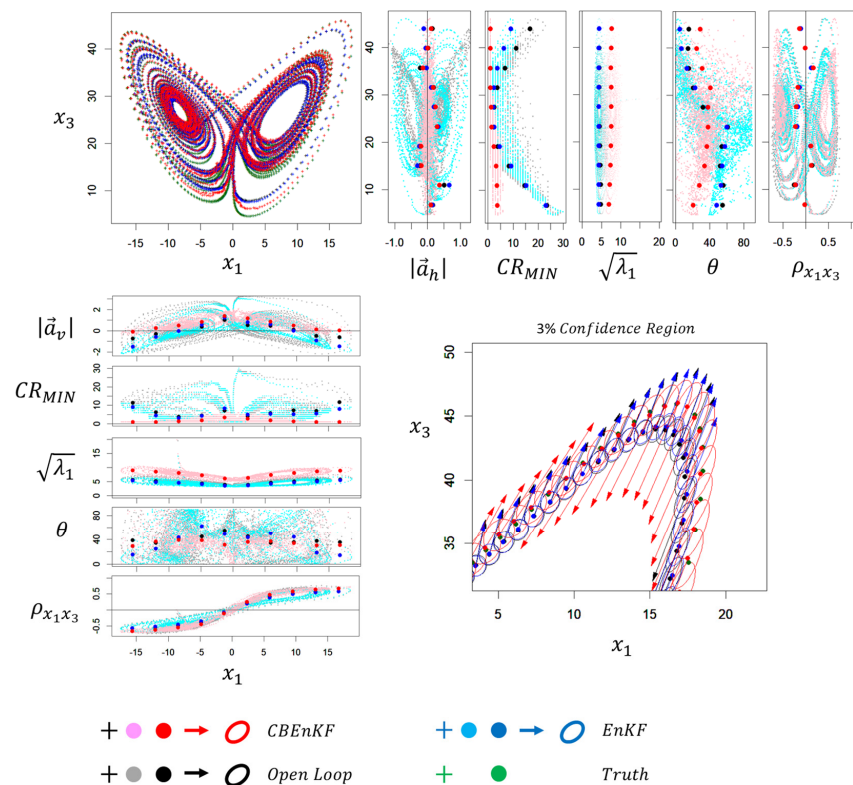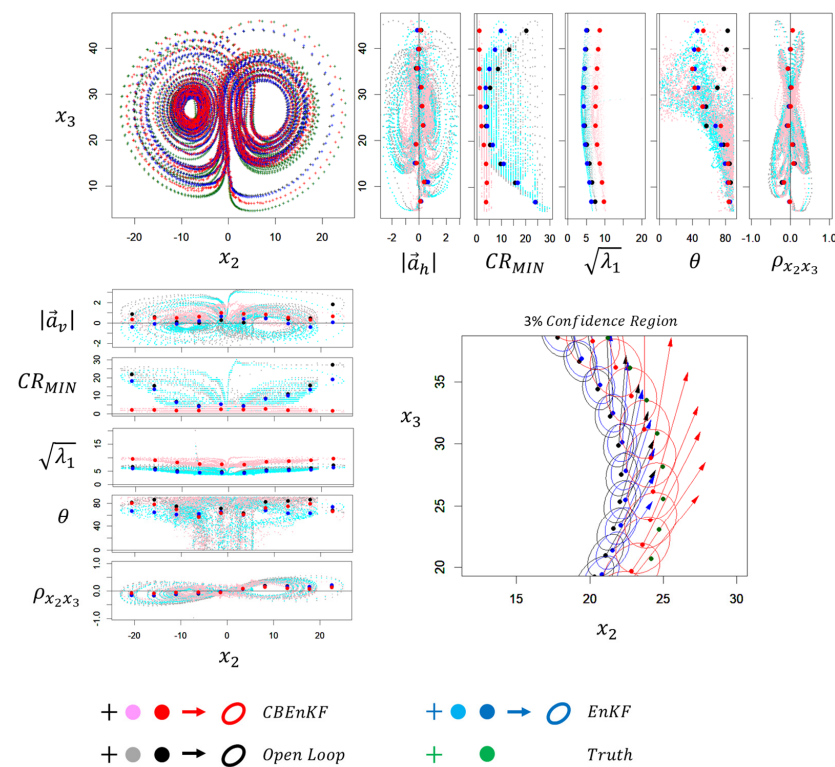


**Figure 7.** Same as Figure 6 but for $x_1$ and $x_3$.

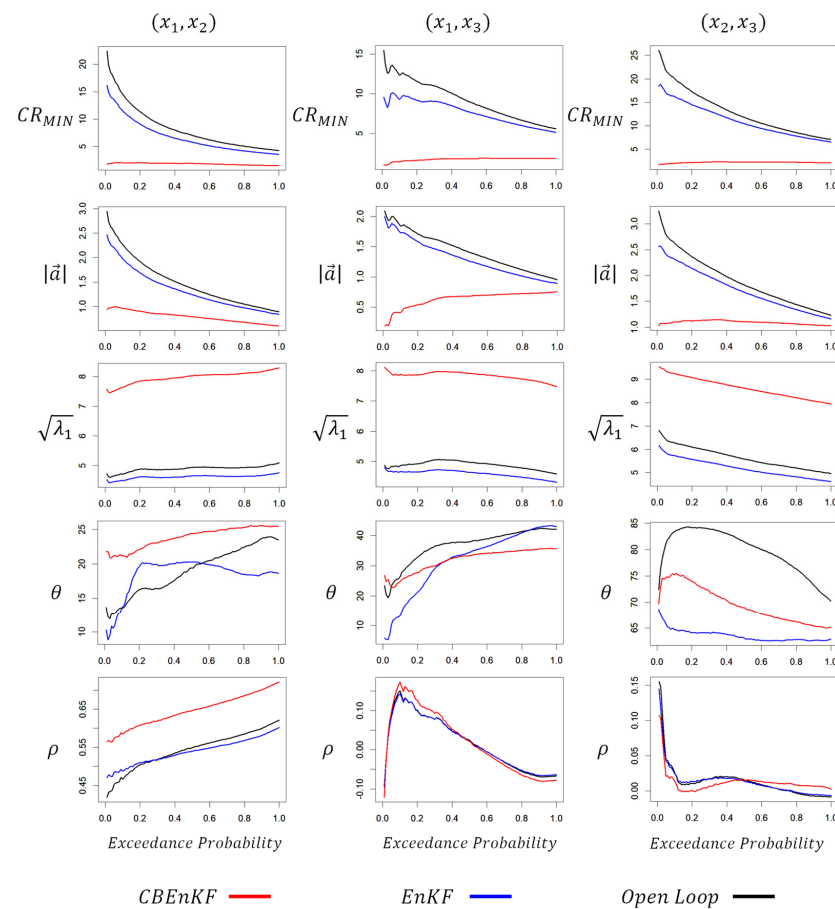**Figure 8.** Same as Figure 6 but for $x_2$ and $x_3$.



**Figure 9.** Geometric statistics of 2-D state-space from the CBEnKF and the EnKF in the case of $n_s = 2000$, and $\sigma_Z^2 = 400$ as a function of an exceedance probability.
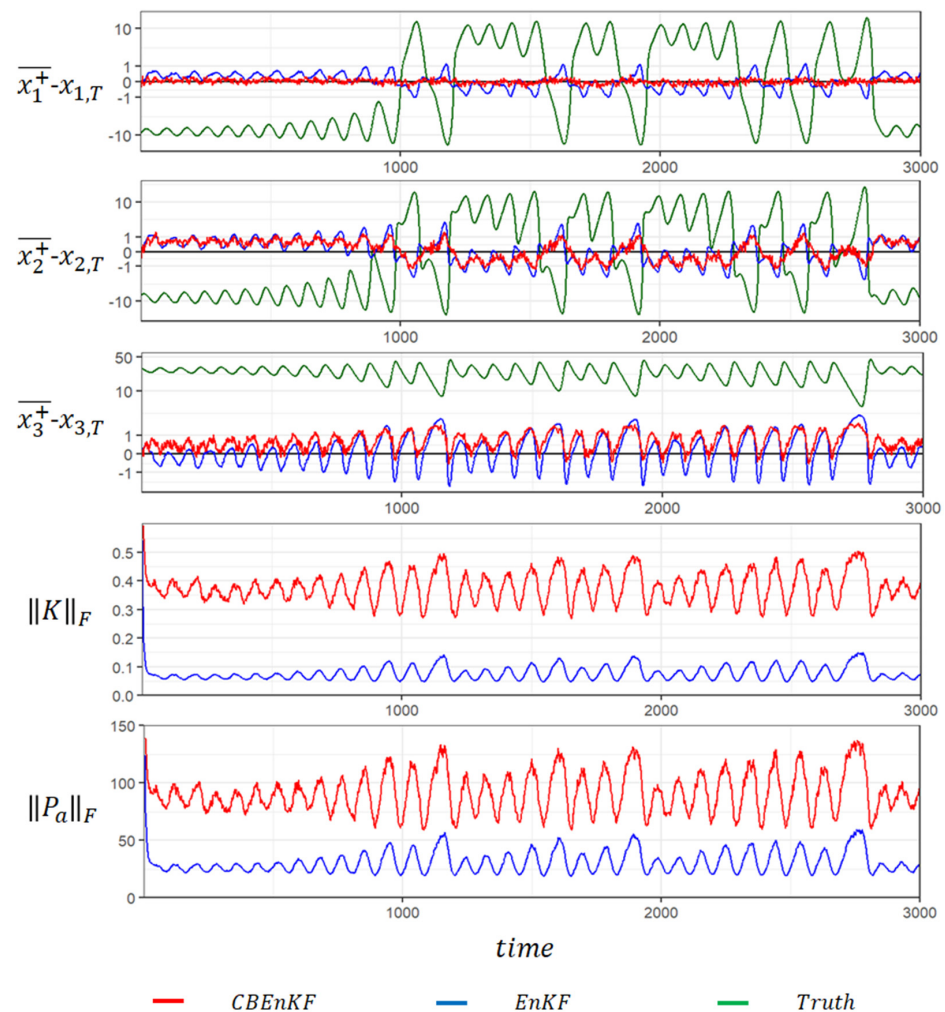
**Figure 10.** Time series of state ensemble mean errors, $||K_k||_F$, and $||P_{a,k}||_F$. At the top three plots, the true state (green line) time series is overlaid for comparison. Inverse Hyperbolic Sine (IHS) transformation is applied to the top three plots to properly view ensemble mean errors, denoted by blue and red lines for the EnKF and the CBEnKF, respectively; $\overline{x}$ denotes the ensemble mean of x; the superscript + represents the updated state; the subscript T denotes the truth; $\overline{||K_{k,i}^C||}_F = 0.38$; $\overline{||K_{k,i}^K||}_F = 0.08$; $\frac{1}{N}\sum_{i=1}^{N} \frac{||K_{k,i}^C||_F}{||K_{k,i}^K||_F} = 5.1$; $\overline{||P_{k,i}^C||}_F = 92$; $\overline{||P_{k,i}^K||}_F = 31$, and $\frac{1}{N}\sum_{i=1}^{N} \frac{||P_{k,i}^C||_F}{||P_{k,i}^K||_F} = 3.1$.
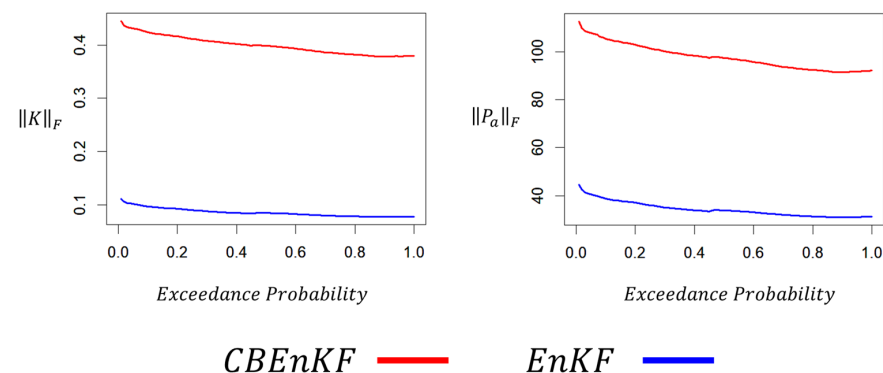


**Figure 11.** Frobenius norm of $K_k$ and $P_{a,k}$, or $||K_k||_F$ and $||P_{a,k}||_F$, respectively, as a function of exceedance probabilities, where $n_s = 2000$, and $\sigma_Z^2 = 400$ are used for both the CBEnKF (red line) and the EnKF (blue line).
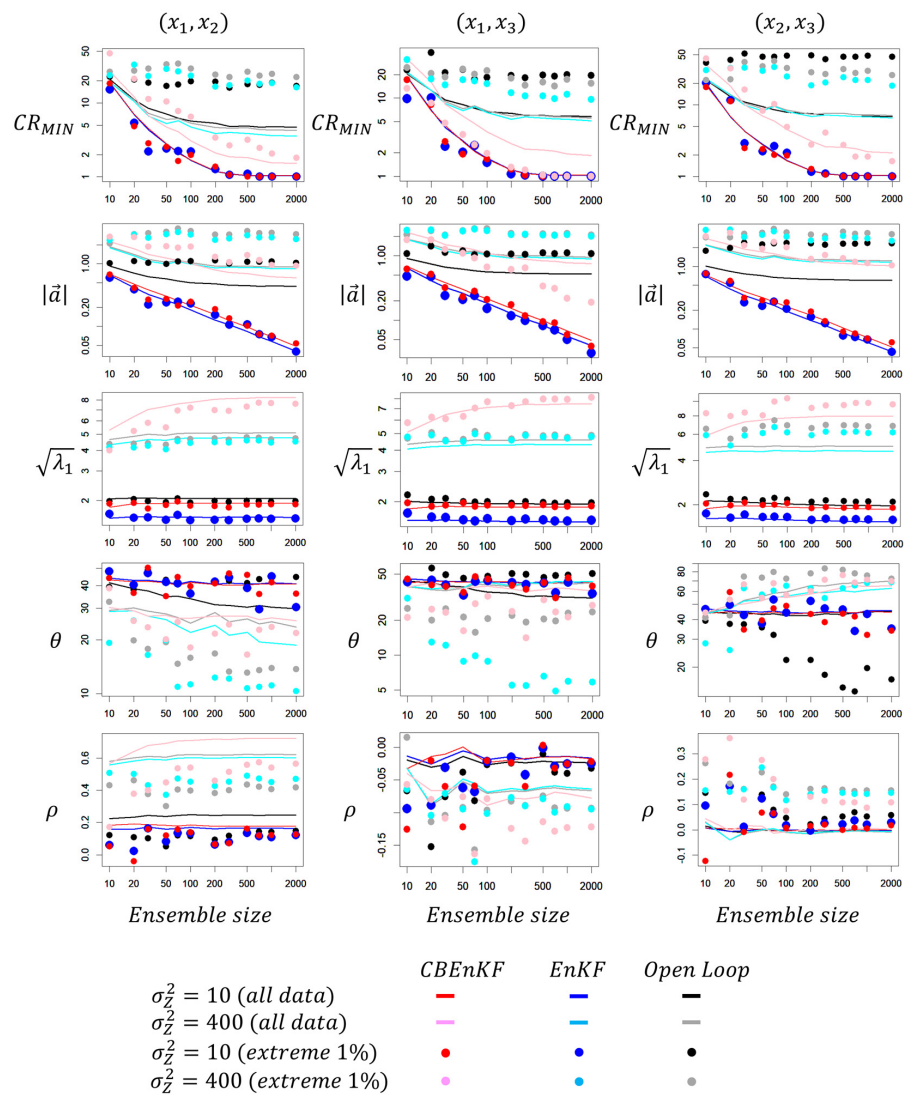
**Figure 12.** Geometric statistics of the 2-D state space from the CBEnKF and the EnKF in the case of $n_s = 10, 20, 30, 50, 70, 100, 200, 300, 500, 700, 1000$, and 2000, and $\sigma_Z^2 = 10$ and 400. Dots (extreme 1%) denote statistics based on samples with the exceedance probability of 0.01.
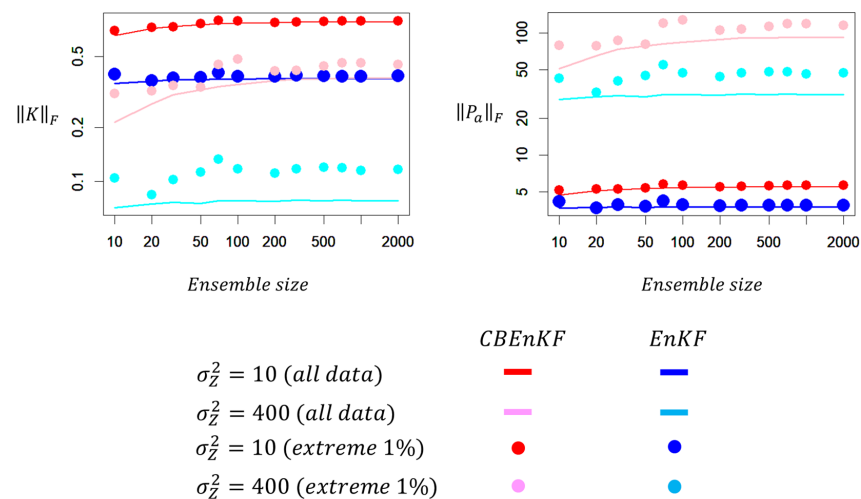


**Figure 13.** Kalman gain norm and analysis covariance norm of the CBEnKF and the EnKF in the case of $n_s = 10, 20, 30, 50, 70, 100, 200, 300, 500, 700, 1000$, and 2000, and $\sigma_Z^2 = 10$ and 400.

Figure 6 presents the results for states $x_1$ and $x_2$ of the Lorenz 63 model in a 2-D state space. The attributes shown include $|\vec{a}_h|$, $|\vec{a}_v|$, $CR_{MIN}$, $\sqrt{\lambda_1}$, $\theta$, $\rho$, and $\sqrt{\lambda_1}\vec{u}_1$; $|\vec{a}_h|$, and $|\vec{a}_v|$ reflect errors in the ensemble mean measured along horizontal or vertical directions, respectively; large dots overlaid in the scatter plots represent mean values of the samples in each of the ten bins equally dividing the entire state space. Figures 7 and 8 are the same as Figure 6 but for the state spaces of $(x_1, x_3)$ and $(x_2, x_3)$, respectively. The following summarizes general observations from Figures 6–8. The spread of $|\vec{a}_h|$ and $|\vec{a}_v|$ in the scatter plots shows that the CBEnKF reduces CBs more effectively than the EnKF or the OL, particularly in the extremes. In some cases, however, their mean values appear similar, e.g., $|\vec{a}_h|$ in the state spaces of $(x_1, x_3)$ or $(x_2, x_3)$. The OL results for $|\vec{a}_h|$ and $|\vec{a}_v|$ show that their patterns appearing in the state space are similar to those of the state space plot in the top left of Figures 6–8; this indicates the notable dependency of the amount of ensemble mean errors on model dynamics, e.g., the larger ensemble mean errors at the extreme—this is also seen in EnKF solutions but less so in the CBEnKF. This may be explained by the CBEnKF with a larger weight to observations than the EnKF in the case of largely uncertain observations ($\sigma_z^2 = 400$), which reduces the reliance of CBEnKF solutions on the model dynamics. Based on $\sqrt{\lambda_1}$, CBEnKF covariances are generally larger than the EnKF at all state spaces. Larger $\sqrt{\lambda_1}$ and smaller $|\vec{a}_h|$ and $|\vec{a}_v|$ of the CBEnKF than the EnKF yield consistently smaller $CR_{MIN}$ than the EnKF at both extremes and the median of all three variables. This signifies the benefit of using the CB-informed KF for the estimation of extremes given that the EnKF's $CR_{MIN}$ quickly increases towards extremes, i.e., the EnKF is less confident in estimating extremes than the CBEnKF. For example, 3% confidence regions for selected extreme values presented in the bottom right plots show the truth (green dots) contained within the CBEnKF's confidence regions (red ellipses) but not within the EnKF's (blue ellipses). At the plots, arrows represent $\sqrt{\lambda_1}\vec{u}_1$. $\theta$ and $\rho$ do not clearly indicate differences between the two filters.

Figure 9 shows $CR_{MIN}$, $|\vec{a}|$, $\sqrt{\lambda_1}$, $\theta$, and $\rho$ of Figures 6–8 but as a function of exceedance probabilities to highlight CB-informed KF performances at extremes. At extremes with low exceedance probabilities, differences between the CBEnKF and the EnKF are vivid in the case of $|\vec{a}|$ and $CR_{MIN}$. On the other hand, $\sqrt{\lambda_1}$ of the CBEnKF is consistently larger than those of the EnKF and the OL across exceedance probabilities. As exceedance probabilities increase, the EnKF's $|\vec{a}|$ becomes similar to the CBEnKF's, implying unconditionally less biased. The CBEnKF keeps consistently low $CR_{MIN}$ at all exceedance probabilities owing to small $|\vec{a}|$ and large $\sqrt{\lambda_1}$, compared to the EnKF or the OL. Since the EnKF seeks orthogonal solutions to minimize analysis covariances, its $\sqrt{\lambda_1}$ is always smaller than the OL's as well as the CBEnKF's. On the other hand, the CBEnKF increases $\sqrt{\lambda_1}$ to address CBs which helps keep $CR_{MIN}$ low to contain the truth. Both $\theta$ and $\rho$ show no consistent patterns across different state spaces as well as exceedance probabilities.

In Figure 10, the ensemble mean error time series indicates that among the three variables, CBEnKF's improvement is the largest for $x_1$. On the other hand, for the state $x_3$, the CBEnKF mainly remedies the underestimation of $x_3$ compared to the EnKF. In the case of $x_2$, the CBEnKF slightly outperforms the EnKF. These observations may imply the different amounts of CBs present in different states, hence the need of applying a separate weight $\alpha$ to the CB penalty for the individual state, which warrants a future effort. To compare $P_{a,k}$ and $K_k$ from the two filters, the time series of Frobenius norm of $P_{a,k}$ and $K_k$ is computed by Equations (56) and (57), respectively. Compared to the EnKF, the CBEnKF yields $||K_k||_F$ and $||P_{a,k}||_F$ consistently larger at all assimilation cycles, and the mean values of $||K_k||_F$ and $||P_{a,k}||_F$ are five and three times larger, respectively.

$$||P_{a,k}||_F = \sqrt{\sum_{i=1}^{3} \lambda_i^2} \tag{56}$$

$$||K_k||_F = \sqrt{\sum_{i=1}^{3} \lambda_i^2} \tag{57}$$

Figure 11 shows mean $||K_k||_F$, and $||P_{a,k}||_F$ as a function of exceedance probabilities. At extremes, both the CBEnKF and the EnKF show that mean $||K_k||_F$ and $||P_{a,k}||_F$ are larger than those at high exceedance probabilities, and that large differences in mean $||K_k||_F$ and $||P_{a,k}||_F$ between the CBEnKF and the EnKF are consistent across exceedance probabilities.

Figures 6–11 are based on the case of uncertain observations ($\sigma_z^2 = 400$) where the CBEnKF may supposedly outperform the EnKF. To explore the CBEnKF performance with less uncertain observations ($\sigma_z^2 = 10$) and also to see the sensitivity to the ensemble size ($n_S$), Figure 12 presents results from the combination of $n_S = 10, 20, 30, 50, 70, 100, 200, 300, 500, 700, 1000,$ and 2000, and $\sigma_z^2 = 10$ and 400. In Figure 12, $|\vec{a}|$ plots indicate that with $\sigma_z^2 = 10$, the accuracy of the ensemble mean continuously increases with an increase of $n_S$ at both cases of extremes (an exceedance probability of 0.1; red and blue dots for the CBEnKF and the EnKF, respectively) and all data (red and blue lines for the CBEnKF and the EnKF, respectively). When $\sigma_z^2 = 10$, the EnKF's $|\vec{a}|$ is slightly smaller than the CBEnKF's, but the CBEnKF's $\sqrt{\lambda_1}$ is slightly larger than the EnKF's. The resulting $CR_{MIN}$ from both filters are very similar. This implies when observations are less uncertain, the EnKF solutions are as accurate and as confident as the CBEnKF solutions at extremes as well as the whole range. When $n_S \geq 200$ and $\sigma_z^2 = 10$, mean $CR_{MIN}$ maintains ~1%. When $n_S < 200$ and $\sigma_z^2 = 10$, $CR_{MIN}$ quickly increases with a decrease of $n_S$ because of inaccurate error covariance estimates with an insufficient ensemble size. When observations are largely uncertain ($\sigma_z^2 = 2000$), the CBEnKF clearly shows more accurate ensemble means (smaller $|\vec{a}|$) and higher confidence in covariance estimates (smaller $CR_{MIN}$) than the EnKF, particularly at extremes. Compared to $\sigma_z^2 = 10$, assimilating largely uncertain observations ($\sigma_z^2 = 2000$) reduces accuracies in covariance estimates, resulting in larger $\sqrt{\lambda_1}$ in both filters, although the CBEnKF's $\sqrt{\lambda_1}$ addressing the CB is larger than the EnKF's. When $\sigma_z^2 = 2000$, $|\vec{a}|$ and $CR_{MIN}$ tend to be less sensitive to $n_S$ than the case of $\sigma_z^2 = 10$. Both $\theta$ and $\rho$ show neither any consistent patterns nor sensitivities to $n_S$, but are included in Figure 12 for completeness.

Finally, Figure 13 presents mean $||K_k||_F$ and $||P_{a,k}||_F$ as a function of $n_S$. Compared to the results from $\sigma_z^2 = 2000$, $\sigma_z^2 = 10$ results in larger $||K_k||_F$ in both filters due to bigger weights to the observations. When $\sigma_z^2 = 2000$, the CBEnKF maintains relatively large $||K_k||_F$ to account for the CB; however, the EnKF's $||K_k||_F$ is conspicuously small. Both $||K_k||_F$ and $||P_{a,k}||_F$ tend to be little sensitive to the ensemble size $n_S$, except the all data case of the CBEnKF with $\sigma_z^2 = 2000$ (pink line). With uncertain observations ($\sigma_z^2 = 2000$), the CBEnKF's $||P_{a,k}||_F$ becomes large at extremes (pink dots) as well as all data (pink line) at all $n_S$ values used to reflect CBs in all states.

## 4. Conclusions

Error covariance and gain matrices of two CB-informed KFs, i.e., the CBPKF and the CBEnKF, are geometrically illustrated and compared with the KF equivalents [19] for a bi-state model using error vectors in the Euclidean space. Geometric illustration and analysis offer an intuitive understanding of the relationship between the two filters. Unlike the KF, the CBPKF solution is not orthogonal to its error, which renders its error covariances and gains to be larger than the KF's. The above differences result in different confidence regions and principal error directions in the state space. Synthetic sensitivity experiments with the Lorenz 63 model showed that the CBEnKF solutions have generally smaller errors in the ensemble mean, larger eigenvalues in the error covariance matrix, more accurate confidence regions for encompassing the truth, and larger Frobenius norms of the error covariance and gain matrices than the KF. The above differences are particularly pronounced when the observations are highly uncertain.

Future research recommendations include applying the CBPKF and the CBEnKF to diverse geophysical problems of estimating and predicting extremes, e.g., extreme precipitation or floods. The bi-state model was used in this work for a comparative geometric analysis of the CBPKF and the KF. Possible extension to an arbitrary number of states poses an interesting research topic.

**Author Contributions:** Conceptualization, H.L., H.S. and D.-J.S.; methodology, H.L. and D.-J.S.; software, H.L. and D.-J.S.; validation, H.L., H.S. and D.-J.S.; formal analysis, H.L.; investigation, H.L., H.S. and D.-J.S.; writing—original draft preparation, H.L.; writing—review and editing, H.L., H.S. and D.-J.S. All authors have read and agreed to the published version of the manuscript.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Nomenclature

| Symbol | Description |
| --- | --- |
| Superscript | |
| C | Conditional bias-penalized Kalman filter (CBPKF) |
| K | Kalman filter (KF) |
| T | Transpose of a matrix |
| $-$ | Forecast |
| Subscript | |
| a | Analysis |
| E | Variable in the Euclidean domain |
| h | Horizontal |
| $k$ | Time step |
| $s$ | Stochastic variable |
| $T$ | Truth |
| $v$ | Vertical |
| $\vec{a}$ | Vector connecting the ensemble mean and the truth |
| $|\vec{a}|$ | Magnitude of the vector $\vec{a}$ |
| C | Pearson product-moment correlation matrix |
| CR | Confidence region |
| $CR_{MIN}$ | Minimum percentage confidence required to contain the verifying truth |
| E | Expected value |
| H | Observation operator that maps model states X to the observation vector Z |
| $\vec{i}$ | Basis vector of the x-axis |
| K | Kalman gain |
| $||K||_F$ | Frobenius norm of K |
| M | The dynamical model |
| n | The total number of observations |
| $n_c$ | The number of variables in the control vector |
| $n_s$ | Ensemble size |
| P | Covariance matrix of state estimate |
| $P^-$ | Covariance matrix of state forecast |
| $P(s < p)$ | Chi-Square probability in the case of $s$ smaller than $p$ |
| Q | The covariance of the dynamical model error W |
| R | The covariance of the observation error vector V |
| s | The value in the Chi-Square probability table for a given confidence region |
| $\vec{u}$ | Eigenvector of the covariance matrix P |
| V | The observation error vector |

| W | The dynamical model error |
|---|---|
| x | Individual model state |
| $\|\vec{x}\|$ | Norm of the vector $\vec{x}$ |
| x* | Realization of an estimated state X* |
| X | Model state vector |
| X* | The vector of the estimated states |
| y | Innovation |
| Z | The observation vector |
| $\alpha$ | The weight given to the CB penalty term in $\Sigma_{EV} + \alpha\Sigma_{CB}$ |
| $\varepsilon$ | Error in a state |
| $\lambda$ | Eigenvalue |
| $\rho_{ij}$ | Correlation between variables i and j |
| $\sigma_{ij}$ | Covariance of variables i and j |
| $\sigma_i^2$ | Variance of a variable i |
| $\sigma_Z^2$ | Observation error variance |
| $\Sigma_{EV}$ | The error variance |
| $\Sigma_{CB}$ | The expectation of the Type-II CB squared |

## Appendix A. Eigenvalue Decomposition of a Two-by-Two Covariance Matrix

The eigenvalue decomposition (EVD) of the covariance matrix $P_{a,k} = \begin{bmatrix} s_{1,k}^2 & s_{12,k} \\ s_{12,k} & s_{2,k}^2 \end{bmatrix}$ can be written as

$$P_{a,k} = \begin{bmatrix} s_{1,k}^2 & s_{12,k} \\ s_{12,k} & s_{2,k}^2 \end{bmatrix} = UEU^T \tag{A1}$$

$$U = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, E = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix},$$

U is the eigenvector matrix that can be interpreted as the rotation matrix applied to white data, or a standard normal and uncorrelated data. *E* is the eigenvalue matrix that explains the variance of the principal error direction, or the direction of the eigenvector. In the case of two-dimensional data, $\sqrt{E}$ can be interpreted as a scale factor applied to white data. In other words, the resulting dataset (D) from scaling white data (W) by $\sqrt{E}$ and then rotating it by *U*, i.e., $D = U\sqrt{E}W$, will have the covariance matrix of $P_{a,k} = U\sqrt{E}\left(U\sqrt{E}\right)^T$.

Let $\lambda_1 > \lambda_2 > 0$, or $P_{a,k}$ is positive definite

$$\begin{aligned} \begin{bmatrix} s_{1,k}^2 & s_{12,k} \\ s_{12,k} & s_{2,k}^2 \end{bmatrix} &= \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \\ &= \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \lambda_1\cos\theta & -\lambda_1\sin\theta \\ \lambda_2\sin\theta & \lambda_2\cos\theta \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1\cos^2\theta + \lambda_2\sin^2\theta & -\lambda_1\sin\theta\cos\theta + \lambda_2\sin\theta\cos\theta \\ -\lambda_1\sin\theta\cos\theta + \lambda_2\sin\theta\cos\theta & \lambda_1\sin^2\theta + \lambda_2\cos^2\theta \end{bmatrix} \end{aligned} \tag{A2}$$

$$s_{1,k}^2 = \lambda_1\cos^2\theta + \lambda_2\sin^2\theta \tag{A3}$$

$$s_{12,k} = (\lambda_2 - \lambda_1)\sin\theta\cos\theta \tag{A4}$$

$$s_{2,k}^2 = \lambda_1\sin^2\theta + \lambda_2\cos^2\theta \tag{A5}$$

A(3) + A(5), A(3) − A(5), and 2 × (A4) yield (A6), (A7), and (A8), respectively.

$$s_{1,k}^2 + s_{2,k}^2 = \lambda_1 + \lambda_2 \tag{A6}$$

$$s_{1,k}^2 - s_{2,k}^2 = \lambda_1\left(\cos^2\theta - \sin^2\theta\right) - \lambda_2\left(\cos^2\theta - \sin^2\theta\right) = (\lambda_1 - \lambda_2)\cos2\theta \tag{A7}$$

$$2s_{12,k} = (\lambda_2 - \lambda_1)2\sin\theta\cos\theta = (\lambda_2 - \lambda_1)\sin2\theta \tag{A8}$$

From (A7) and (A8)

$$(\lambda_1 - \lambda_2)^2 = \left(s_{1,k}^2 - s_{2,k}^2\right)^2 + 4s_{12,k}^2 \tag{A9}$$

Since $\lambda_1 > \lambda_2$,

$$\lambda_1 - \lambda_2 = \sqrt{\left(s_{1,k}^2 - s_{2,k}^2\right)^2 + 4s_{12,k}^2} \tag{A10}$$

(A6) + (A10)

$$\lambda_1 = \frac{1}{2}\left(s_{1,k}^2 + s_{2,k}^2 + \sqrt{\left(s_{1,k}^2 - s_{2,k}^2\right)^2 + 4s_{12,k}^2}\right) \tag{A11}$$

$$\lambda_2 = \frac{1}{2}\left(s_{1,k}^2 + s_{2,k}^2 - \sqrt{\left(s_{1,k}^2 - s_{2,k}^2\right)^2 + 4s_{12,k}^2}\right) \tag{A12}$$

From (A7) and (A8)

$$\tan 2\theta = -\frac{2s_{12,k}}{s_{1,k}^2 - s_{2,k}^2} \tag{A13}$$

$$\theta = \frac{1}{2}\tan^{-1}\left(-\frac{2s_{12,k}}{s_{1,k}^2 - s_{2,k}^2}\right) \tag{A14}$$

## References

1. Kalman, R.E. A new approach to linear filtering and prediction problems. *J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
2. Evensen, G. The Ensemble Kalman Filter: Theoretical formulation and practical implementation. *Ocean Dyn.* **2003**, *53*, 343–367. [CrossRef]
3. Seo, D.-J.; Saifuddin, M.M.; Lee, H. Conditional bias-penalized Kalman filter for improved estimation and prediction of extremes. *Stoch. Environ. Res. Risk Assess.* **2018**, *32*, 183–201. [CrossRef]
4. Seo, D.-J.; Saifuddin, M.M.; Lee, H. Correction to: Conditional bias-penalized Kalman filter for improved estimation and prediction of extremes. *Stoch. Environ. Res. Risk Assess.* **2018**, *32*, 3561–3562. [CrossRef]
5. Lee, H.; Shen, H.; Noh, S.J.; Kim, S.; Seo, D.-J.; Zhang, Y. Improving flood forecasting using conditional bias-penalized ensemble Kalman filter. *J. Hydrol.* **2019**, *575*, 596–611. [CrossRef]
6. Ciach, G.J.; Morrissey, M.L.; Krajewski, W.F. Conditional bias in radar rainfall estimation. *J. Appl. Meteorol.* **2000**, *39*, 1941–1946. [CrossRef]
7. Fuller, W.A. *Measurement Error Models*; Wiley: Chichester, UK, 1987.
8. Seber, G.A.F.; Wild, C.J. *Nonlinear Regression*; John Wiley & Sons: New York, NY, USA, 1989.
9. Hausman, J.A. Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *J. Econ. Perspect.* **2001**, *15*, 57–67. [CrossRef]
10. Hughes, M.D. Regression dilution in the proportional hazards model. *Biometrics* **1993**, *49*, 1056–1066. [CrossRef] [PubMed]
11. Frost, C.; Thompson, S. Correcting for regression dilution bias: Comparison of methods for a single predictor variable. *J. R. Stat. Soc. Ser. A* **2000**, *163*, 173–190. [CrossRef]
12. Joliffe, I.T.; Stephenson, D.B. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*; Wiley: New York, NY, USA, 2003.
13. Kim, B.; Seo, D.-J.; Noh, S.; Prat, O.P.; Nelson, B.R. Improving multisensor estimation of heavy-to-extreme precipitation via conditional bias-penalized optimal estimation. *J. Hydrol.* **2018**, *556*, 1096–1109. [CrossRef]
14. Wilks, D.S. *Statistical Methods in the Atmospheric Sciences*, 2nd ed.; Academic Press: London, UK, 2006; p. 648.
15. Brown, J.D.; Seo, D.-J. Evaluation of a nonparametric post-processor for bias correction and uncertainty estimation of hydrologic predictions. *Hydrol. Processes* **2013**, *27*, 83–105. [CrossRef]
16. Seo, D.-J.; Siddique, R.; Zhang, Y.; Kim, D. Improving real-time estimation of heavy-to-extreme precipitation using rain gauge data via conditional bias-penalized optimal estimation. *J. Hydrol.* **2014**, *519 Pt B*, 1824–1835. [CrossRef]
17. Shen, H.; Lee, H.; Seo, D.-J. Adaptive conditional bias-penalized Kalman filter for improved estimation of extremes and its approximation for reduced computation. *Hydrology* **2022**, *9*, 35. [CrossRef]
18. Shen, H.; Seo, D.-J.; Lee, H.; Liu, Y.; Noh, S. Improving flood forecasting using conditional bias-aware assimilation of streamflow observations and dynamic assessment of flow-dependent information content. *J. Hydrol.* **2022**, *605*, 127247. [CrossRef]
19. Kronhamn, T.R. Geometric illustration of the Kalman filter gain and covariance update algorithms. *IEEE Control Syst. Mag.* **1985**, *5*, 41–43. [CrossRef]
20. Lorentzen, R.J.; Nævdal, G. An iterative ensemble Kalman filter. *IEEE Trans. Autom. Control* **2011**, *56*, 5–8. [CrossRef]

21.  Houtekamer, P.L.; Zhang, F. Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.* **2016**, *144*, 4489–4532. [CrossRef]
22.  Astrom, K.J. *Introduction to Stochastic Control Theory*; Academic Press: New York, NY, USA, 1970.
23.  Schweppe, F.C. *Uncertain Dynamic Systems*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1973; p. 563.
24.  Kreyszig, E. *Advanced Engineering Mathematics*, 10th ed.; Wiley: New York, NY, USA, 2011.
25.  Lorenz, E.N. Deterministic nonperiodic flow. *J. Atmos. Sci.* **1963**, *20*, 130–141. [CrossRef]
26.  Shen, H.; Lee, H.; Noh, S.; Kim, S.; Seo, D.-J.; Zhang, Y. Conditional bias-penalized Kalman filter for improved state estimation over the tails of distribution. In Proceedings of the AGU Fall Meeting, Washington, DC, USA, 10–14 December 2018.