

集合变换卡尔曼滤波数据同化方法中的协方差局地化

王际朝¹, 王 玥¹, 臧绍东¹, 杨俊钢², 纪艳菊¹, 阮宗利¹

(1. 中国石油大学(华东) 理学院, 山东 青岛 266580; 2. 自然资源部第一海洋研究所, 山东 青岛 266061)

摘要: 在集合数据同化中, 协方差局地化(covariance localization, CL)方法的使用存在限制。集合变换卡尔曼滤波(ensemble transform Kalman filter, ETKF)作为集合平方根滤波的变种方法, 是一种应用较广、计算高效的数据同化方法。本文分析了 CL 方法应用于 ETKF 方法的困难, 从而改进 CL 方法使其可以适用于 ETKF 方法。另外, 结合浅水方程, 利用 Askey 函数作为多元局地化函数, 提出了一种适用于多元数值模型的 CL 方法。通过具体实验验证, 得到了较好的分析结果。

关键词: 协方差局地化; 集合变换卡尔曼滤波; 多元局地化函数

中图分类号: P732 文献标识码: A 文章编号: 1000-3096(2021)06-0034-10

DOI: 10.11759/hyxx20200321001

数据同化是通过对观测数据与预测值进行分析, 寻找一个当前物理状态的最优估计值(分析值), 作为数值预测模型初始条件的方法。作为连接数值模式和观测数据的技术手段, 数据同化在海洋科学领域得到了广泛关注。当前, 集合数据同化方法得以快速发展, 已经成为业务化数值预测的一个可行选择。集合卡尔曼滤波(ensemble Kalman filter, EnKF)^[1], 可以将卡尔曼滤波应用到强非线性模型, 扩展了卡尔曼滤波的适用性, 是数据同化领域的研究热点之一。EnKF 中的预测误差协方差矩阵通过对预测集合成员进行统计计算得到, 预测误差对于集合数据同化方法而言至关重要。然而, 产生足够大的预测集合的计算成本令人望而却步。当使用的集合数目过小时, 则不能在统计意义上充分表示模型的变化, 从而会引起欠采样(under-sampling)。通常, 欠采样会导致滤波发散、预测误差协方差低估和虚假相关等负面问题。由于这些问题的影响, EnKF 可能会产生一个次优的分析结果。

在进行数据同化时, 为了克服欠采样带来的负面影响, 有关学者已经提出了很多解决方法。当前基于集合的数据同化研究中, 最著名的解决方法是协方差膨胀(covariance inflation, CI)^[2]、协方差局地化(covariance localization, CL)^[3-4]和局地化分析(local analysis, LA)^[5-6]。CI 方法通过一个膨胀因子校正预测误差协方差的低估问题, 主要思想是通过膨胀集合均值和集合成员之间的误差来增大预测误差协方差。CI 方法用于分析过程之前, 具体表示

为: $x_i^f \leftarrow r(x_i^f - \bar{x}^f) + \bar{x}^f$, 其中“ \leftarrow ”表示对于之前状态值的替代, r 是所谓的膨胀因子。尽管 CI 方法克服了预测误差协方差的低估问题, 但是协方差膨胀因子 r 解决不了虚假相关问题。CL 方法在预测误差协方差矩阵和局地化函数之间实施舒尔积运算^[7], 通过截断预测误差协方差矩阵中预先指定距离之外的误差相关性以解决虚假相关问题。此外, 舒尔积运算能够提高预测误差协方差矩阵的秩^[8]。LA 方法以待更新状态变量为中心, 建立一个虚拟的局部窗口, 得到该状态变量预测误差协方差的局部近似值, 在分析更新过程中, 将局部窗口外的集合扰动设定为零。例如, 通常可以以某个状态变量为中心, 仅仅同化与其固定距离为 M 之内的观测数据, 固定距离的长度决定了同化过程中观测值的数量。显然, 分别对状态变量进行局部同化的特点使 LA 方法可以利用并行方法进行快速计算。

收稿日期: 2020-03-21; 修回日期: 2020-06-11

基金项目: 国家重点研发计划项目(2018YFC1406200, 2016YFC1401800); 国家自然科学基金项目(41406007); 中央高校基本科研业务费专项资金(19CX05003A-5); 山东省自然科学基金面上项目(ZR2020MD060)

[Foundation: Foundation of the National Key Research and Development Project of China, No. 2018YFC1406200, No. 2016YFC1401800; National Natural Science Foundation of China, No. 41406007; Fundamental Research Funds for the Central Universities, No. 19CX05003A-5; Shandong Provincial Natural Science Foundation, No. ZR2020MD060]

作者简介: 王际朝(1979—), 男, 山东莘县人, 副教授, 博士, 主要从事海洋数值模拟与数据同化研究, 电话: 0532-86983537, E-mail: wangjc@upc.edu.cn

当前,越来越多的学者^[9-11]关注的是CL方法和LA方法。相比于CL方法,LA方法是一种独立的方法,它可以用于任何数据同化的框架,但是它的实际同化效果并不如CL方法。Miyoshi等^[12]指出LA方法的同化效果和CL方法类似,但是LA方法通常会导致弱局地化影响。之后,Janjić等^[9]通过对Lorenz96模型进行实验发现:如果观测误差方差小于初始状态的预测误差方差,CL方法将会导致比较小的估计误差;如果观测误差方差大于初始状态的预测误差方差,CL方法和LA方法有相同的局地化影响。因此,从之前的研究和实验可以得出,CL方法可能是相对较好的一个局地化方法。由于CL方法中的舒尔积运算仅仅用于预测误差协方差矩阵,而集合转换卡尔曼滤波(ensemble transform Kalman filter, ETKF)方法中的预测误差协方差矩阵并没有显式计算,而是传递预测集合扰动矩阵,这导致ETKF方法中进行舒尔积运算的矩阵维度不一致,所以CL方法不能用于ETKF方法^[13]。Janjić等^[9]的文章给出了CL方法不适用于ETKF方法的具体讨论。

为了克服CL方法不能用于ETKF方法的限制,在一元模型中,通过对局地化函数的平方根矩阵和预测集合扰动矩阵进行舒尔积运算, Petrie^[14]提出了一种用于ETKF方法的CL近似方法。对于矩阵维度不一致的问题,通过在预测集合扰动矩阵中添加 $n-N$ 列零向量(n 表示状态变量个数, N 表示集合个数),对矩阵进行扩展来修正。然而,韩培等^[15]对此方法进行探究,表明这种近似的CL方法是一种弱近似并产生了较差的同化效果。

基于集合样本相关性和矩阵因式分解,近几年提出了动态计算局地化函数的方法,使得CL方法可以应用到ETKF方法中。Bishop等^[16]提出了一种生成局地化函数的新方法,该函数能够随真实误差相关函数移动并且还适应真实误差相关函数的宽度。但是此方法的计算代价比较大,鉴于此,Bishop等^[17]提出了一种新的协方差自适应局地化方法(CALECO)的具体实施方案,成功地将CL方法应用到ETKF方法,但是此方法仍然具有较高的计算复杂度。基于上述的研究基础,Bishop等^[18]在ETKF方法中提出了较为可行的CL方法的实施方案,并命名为增益集合转换卡尔曼滤波(gain form of ensemble transform Kalman filter, GETKF)。GETKF中使用的是Gaspari等^[19]提出的基于距离相关的一个

局地裁剪函数作为局地化函数(GC局地化函数),此方法解决了在一元数值预测模型中CL方法不能用于ETKF方法的问题。

当前,由于CL方法的实现存在限制,国内外对于CL方法在ETKF等平方根滤波中的研究较少。本文将基于GETKF方法进行研究,提出一种局地化效果较好的局地化方法,并通过数值模拟实验比较本文基于GETKF修改的方法与GETKF方法的同化效果。

1 理论背景与方法改进

Campbell等^[20]比较了EnKF方法中分别使用模型空间CL方法和观测空间CL方法的同化效果,认为模型空间CL方法通常优于观测空间CL方法。因此,本文以下的研究基于模型空间CL方法。本节首先介绍GETKF方法,然后对该方法进行改进。

1.1 增益集合转换卡尔曼滤波

在ETKF方法中,应用CL方法的难点在于预测误差协方差矩阵并没有显式表示,而是通过预测集合扰动矩阵隐式表达,因而在预测集合扰动矩阵和局地化函数之间无法进行舒尔积运算,当前主要的目的是在预测集合扰动矩阵和局地化函数的平方根矩阵中进行舒尔积运算来近似预测误差协方差矩阵和局地化函数之间的舒尔积运算。假定预测集合扰动矩阵表示为

$$\begin{aligned} \mathbf{X}'^f &= \frac{1}{\sqrt{N-1}}(x_1^f - \bar{x}^f, x_2^f - \bar{x}^f, \dots, x_N^f - \bar{x}^f), \\ &= (u_1, u_2, \dots, u_N), \end{aligned} \quad (1)$$

预测误差协方差矩阵为 \mathbf{P}^f ,则有 $\mathbf{P}^f = \mathbf{X}'^f (\mathbf{X}'^f)^T$ 。

局地化函数 ρ 的平方根矩阵表示为 \mathbf{W} ,即

$$\rho = \mathbf{W}\mathbf{W}^T = (w_1, w_2, \dots, w_L)(w_1, w_2, \dots, w_L)^T, \quad (2)$$

其中,平方根矩阵 \mathbf{W} 通过对局地化函数 ρ 进行特征值分解得到,并且按特征值大小降序排列,取前10个特征值和特征向量对构成,即

$$\mathbf{W} = \rho_{\text{eigenvectors}} (\rho_{\text{eigenvalue}})^{1/2}. \quad (3)$$

一般情况,CL方法需要进行如下形式舒尔积运算,

$$\mathbf{P}_{\text{loc}}^f = \mathbf{P}^f \circ \rho. \quad (4)$$

由于之前讨论的ETKF方法的局限性,GETKF方法中采用了集合扩展技术定义一个 $n \times (N \times L)$ 维度的矩阵 \mathbf{Z}^f ,

$$\mathbf{Z}^f = \mathbf{W} \Delta \mathbf{X}'^f = \left[(w_1 \circ u_1, w_1 \circ u_2, \dots, w_1 \circ u_N), \dots, (w_L \circ u_1, w_L \circ u_2, \dots, w_L \circ u_N) \right]. \quad (5)$$

定义矩阵 \mathbf{Z}^f 表示局地化后的预测误差协方差矩阵 $\mathbf{P}_{\text{loc}}^f$ 的平方根矩阵, 即

$$\mathbf{P}_{\text{loc}}^f = \mathbf{P}^f \circ \rho = \mathbf{Z}^f (\mathbf{Z}^f)^T. \quad (6)$$

Bishop 等^[18]已经证明式(6)是成立的。因此成功地在预测集合扰动矩阵 \mathbf{X}'^f 实现了 CL 方法。现在使用 \mathbf{Z}^f 代替 \mathbf{X}'^f 作为当前的预测集合扰动矩阵, 即局地化后的预测误差协方差矩阵 $\mathbf{P}_{\text{loc}}^f$ 的平方根矩阵, 重新构造集合扩展后的预测集合和预测误差协方差矩阵。

令 $M = N \times L$, 则扩展预测集合表示为 $V^f = (v_1^f, v_2^f, \dots, v_M^f)$, 其中每个列向量 v_k^f 如下定义:

$$v_k^f = \frac{1}{N} \sum_{j=1}^N x_j^f + \sqrt{M} z_k, \quad k = 1, 2, \dots, M, \quad (7)$$

其中, z_k 表示矩阵 \mathbf{Z}^f 的第 k 列。很显然扩展预测集

$$\mathbf{Y}'^f = \mathbf{H} \mathbf{Z}^f = (M)^{-1/2} \left[\mathbf{H}(v_1^f) - \mathbf{H}(\bar{v}^f), \mathbf{H}(v_2^f) - \mathbf{H}(\bar{v}^f), \dots, \mathbf{H}(v_M^f) - \mathbf{H}(\bar{v}^f) \right], \quad (10)$$

$$\mathbf{H}(\bar{v}^f) = \frac{1}{M} \sum_{i=1}^M \mathbf{H}(v_i^f). \quad (11)$$

矩阵 \mathbf{H} 表示观测算子。为了提高 ETKF 方法同化的计算效率, 对标准化的预测-观测集合扰动矩阵 $\hat{\mathbf{Y}}^f$ 进行奇异值分解:

$$\hat{\mathbf{Y}}^f = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T, \quad (12)$$

其中, 矩阵 \mathbf{U} 和 \mathbf{V} 是正交矩阵, $\hat{\mathbf{Y}}^f$ 的奇异值由矩阵 $\mathbf{\Sigma}$ 的对角元素给出。接下来, 分析误差协方差矩阵的平方根矩阵(分析集合扰动矩阵) \mathbf{Z}^a 表示为,

$$\mathbf{Z}^a = \mathbf{Z}^f \mathbf{U} (\mathbf{\Sigma} \mathbf{\Sigma}^T + \mathbf{I})^{-\frac{1}{2}} \mathbf{U}^T. \quad (13)$$

分析集合均值 \bar{v}^a 和分析集合 V^a 可以分别表示如下:

$$\begin{aligned} X_{\text{raw}}^a &= \sqrt{N-1} \mathbf{X}'^f - \mathbf{Z}^f \mathbf{U} \left[\mathbf{I} - (\mathbf{S} \mathbf{S}^T + \mathbf{I})^{-\frac{1}{2}} \right] (\mathbf{S} \mathbf{S}^T)^{-1} \mathbf{U}^T (\hat{\mathbf{Y}}^f)^T \mathbf{R}^{-\frac{1}{2}} \mathbf{H} \sqrt{N-1} \mathbf{X}'^f, \\ \alpha &= \left\{ \frac{\text{trace} \left[\mathbf{Z}^a (\mathbf{Z}^a)^T \right]}{\text{trace} \left[\mathbf{X}_{\text{raw}}^a (\mathbf{X}_{\text{raw}}^a)^T / \sqrt{N-1} \right]} \right\}^{1/2}. \end{aligned} \quad (17)$$

这样, 便得到了与原始预测集合大小一致的分析集合 X^a , 然后带入系统动力学模型进行下一时刻预测即可。

以上展示了 GETKF 方法的主要计算过程, 可以看出此方法存在改进的空间: 首先, GETKF 方法

合的均值与原始预测集合均值相同。扩展预测集合误差协方差矩阵表示为:

$$\mathbf{P}_{\text{loc}}^f = \frac{1}{M} \sum_{i=1}^M (v_i^f - \bar{v}^f) (v_i^f - \bar{v}^f)^T. \quad (8)$$

GETKF 方法的分析过程与标准的 ETKF 方法的分析过程类似, 区别是 GETKF 方法使用扩展预测集合 $\{v_i^f, i=1, 2, \dots, M\}$ 以及扩展预测集合扰动矩阵 \mathbf{Z}^f 代替原始预测集合 $\{x_i^f, i=1, 2, \dots, N\}$ 和原始预测集合扰动矩阵 \mathbf{X}'^f 。

定义标准化的预测-观测集合扰动矩阵 $\hat{\mathbf{Y}}^f$ 如公式(7)所示:

$$\hat{\mathbf{Y}}^f = \mathbf{R}^{-\frac{1}{2}} \mathbf{Y}'^f, \quad (9)$$

其中,

$$\bar{v}^a = \bar{v}^f + \mathbf{Z}^f \mathbf{U} \mathbf{\Sigma} (\mathbf{\Sigma}^T \mathbf{\Sigma} + \mathbf{I})^{-1} \mathbf{V}^T \mathbf{R}^{-\frac{1}{2}} \left[\mathbf{y} - \mathbf{H}(\bar{v}^f) \right], \quad (14)$$

$$V^a = \bar{v}^a \mathbf{1}_M^T + \sqrt{M} \mathbf{Z}^a. \quad (15)$$

其中, $\mathbf{1}_M^T = (1, 1, \dots, 1)$ 表示一个元素全为 1 的 $1 \times M$ 维的行向量。很显然, 分析集合的维度是 $n \times M$, 即当前产生了 M 列扩展分析集合, 这是由于我们分析过程使用的预测集合是通过集合扩展后得到的预测集合。但是预测过程仅仅能够传递 N 列分析集合, 因此需要将分析集合的个数 M 降到原始预测集合大小。GETKF 方法中通过构造一个膨胀因子 α , 利用原始预测集合和扩展后的预测集合之间的关系, 构造了 N 列分析集合,

$$X^a = \bar{v}^a \mathbf{1}^T + \alpha X_{\text{raw}}^a, \quad (16)$$

其中,

利用集合扩展技术在分析过程中产生 M 个集合 ($N < M$), 然后使用集合扩展之后的分析误差协方差矩阵计算 N 个集合的方法较为复杂, 如何更加简便地将 M 个集合转换为 N 个集合进行预测过程是值得商榷的; 其次, GETKF 方法中计算扩展后的集合误差协方差矩阵采用有偏形式, 本文对此做出了修改, 采用无偏形式的矩阵进行计算; 此外, GETKF 方法中使用的 GC 局地化函数是一个基于距离相关的局地裁剪函数, 该函数是一元局地化函数, 所以扩展

到多元模型存在限制;最后,由于 GETKF 方法将参数 L 固定为 10,所以在系统状态变量个数较大的情况下,扩展后的预测集合扰动矩阵不能很好地表示误差变化。可以看出 GETKF 方法确实存在的不足之处,对于 GETKF 方法的具体改进将在下一节进行说明。

1.2 增益集合转换卡尔曼滤波的改进

基于 GETKF 方法,本节从局地化函数平方根矩阵的选取、预测误差协方差的无偏估计、随机子采样方法和 GC 局地化函数的限制 4 个方面对原方法进行改进,以提高同化方法的效果。

1.2.1 局地化函数平方根的选取

对于公式(3)中局地化函数平方根矩阵列数 L 的选取,GETKF 方法按照特征值降序排列,取前十个特征值和对应的特征向量构成平方根矩阵 W ,即固定地取 L 为 10。很显然,在应对状态变量个数 n 很大时,前 10 个特征值和特征向量不能完全表示系统变量的误差变化,所以本文重新定义了 L 的选取:

$$L = [\min(10, E_{10\%}), \max(10, E_{10\%})], \quad (19)$$

其中, $E_{10\%}$ 表示前 10% 的特征值的数量。即给定 L 的范围,通过进行多次数值实验选取最优的数值 L 。

1.2.2 预测误差协方差的无偏估计

通常,我们假定卡尔曼滤波中的误差协方差矩阵是无偏的。在统计学上,通过对集合成员与集合均值的偏差除以集合自由度(样本个数减一)来实现。但是 GETKF 方法中的公式(5)~(7)采用的是有偏形式的误差协方差矩阵,这导致误差的产生,造成对误差协方差的低估,所以本文通过将公式(5)~(7)中的 \sqrt{M} 改为 $\sqrt{M-1}$ 来修正协方差公式,构造无偏形式的误差协方差矩阵,从而在进行分析时减少对同化结果的影响。

1.2.3 随机子采样方法

由于扩展后的分析集合数量 M 大于原始预测集合数量 N ,导致集合数量不能在预测模型中进行传递。GETKF 方法中利用一个膨胀因子[式(12)]来克服此问题,该方法需要计算扩展后的分析误差协方差矩阵以及原始预测误差协方差矩阵,如果变量的个数 n 很大,计算难度大大增加。本文使用随机子采样方法进行 N 列分析集合的选取,计算简便。随机子采样方法如下:

$$X^a = \bar{v}^a \mathbf{1}_N^T + \sqrt{N-1} \mathbf{Z}^a \mathbf{randn}(M, N), \quad (20)$$

其中, $\mathbf{1}_N^T = (1, 1, \dots, 1)$ 表示一个元素全为 1 的 $1 \times N$ 维的行向量。式中 $\mathbf{randn}(M, N)$ 表示一个元素服从标准正态分布的维度为 $M \times N$ 的随机矩阵,能够将矩阵 \mathbf{Z}^a 随机地转换为一个维度为 $n \times N$ 的矩阵,从而完成对 N 列分析集合的选取。

1.2.4 GC 局地化函数的限制

GETKF 方法中使用 GC 函数作为局地化函数,但是 GC 函数仅仅是一个一元变量的局地化函数,多元模型中不同变量之间的相关性应该是有区别的,GC 函数并没有体现这一性质,所以在多元模型中使用 GETKF 方法存在限制。本文选用一个基于距离的多元函数 Askey 函数^[21]作为局地化函数,当前此函数已经用于 EnKF 方法,并且在多元模型中进行了实验验证,得到了较好的同化效果。利用 Askey 函数将本文修改的方法扩展到多元模型变量中,探究方法在多元情况下的适用性。

2 实验方案

2.1 模型

2.1.1 KS 模型

KS 模型的表达式为,

$$\frac{\partial u}{\partial t} = -u \frac{\partial u}{\partial x} - \frac{\partial u^2}{\partial x} - \frac{\partial^4 u}{\partial x^4}. \quad (21)$$

KS 模型是一个一元无量纲的非线性偏微分方程模型,方程中二阶项和四阶项会增加模型的复杂性并导致混沌行为。在本文中,KS 模型的偏微分方程通过一个指数时间的四阶龙格库塔数值格式进行离散,时间步为 $\Delta t = 0.25$,终止时间 T 为 250。

2.1.2 浅水模型

浅水模型作为一个多元模型,在当前的数据同化研究中广泛使用。其忽略摩擦效应、科氏力以及非线性项的方程表示为:

$$\frac{\partial h}{\partial t} = -\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}, \quad (22)$$

$$\frac{\partial u}{\partial t} = -\frac{\partial \left(u^2 h + \frac{1}{2} g h^2 \right)}{\partial x} - \frac{\partial (u v h)}{\partial y}, \quad (23)$$

$$\frac{\partial v}{\partial t} = -\frac{\partial (u v h)}{\partial x} - \frac{\partial \left(v^2 h + \frac{1}{2} g h^2 \right)}{\partial y}. \quad (24)$$

其中, u 和 v 分别表示水平和垂直速度, 水面高度由 h 表示, g 为取值为 9.8 m/s^2 的重力加速度。模型区域选择矩形区域 $[0, L] \times [0, L]$, L 取 200 km 。浅水模型的偏微分方程组使用 Lax-Wendroff 有限差分方法计

$$\rho = \begin{cases} 1 - \frac{5}{3} \left(\frac{|z|}{c}\right)^2 + \frac{5}{8} \left(\frac{|z|}{c}\right)^3 + \frac{1}{2} \left(\frac{|z|}{c}\right)^4 - \frac{1}{4} \left(\frac{|z|}{c}\right)^5, & 0 \leq |z| \leq c \\ 4 - \frac{2}{3} \left(\frac{|z|}{c}\right)^{-1} - 5 \left(\frac{|z|}{c}\right) + \frac{5}{3} \left(\frac{|z|}{c}\right)^2 + \frac{5}{8} \left(\frac{|z|}{c}\right)^3 - \frac{1}{2} \left(\frac{|z|}{c}\right)^4 + \frac{1}{12} \left(\frac{|z|}{c}\right)^5, & c \leq |z| \leq 2c \\ 0, & 2c \leq |z| \end{cases} \quad (25)$$

式中, z 表示网格点之间或网格点与观测点之间的距离。由于本文使用模型空间 CL 方法, 所以 z 表示物理空间中网格点之间的距离。局地化长度尺度 c 定义为 $c = \sqrt{10/3}r$, r 为局地化半径, 是一个可选

$$\rho_{i,j}(z; v, c) = \begin{cases} c^{v+1} B(u_{ij} + 1, v + 1) \left(1 - \frac{|z|}{c}\right)^{v+u_{ij}+1}, & |z| < c, \quad i, j = 1, 2, \dots, m \\ 0, & |z| \geq c \end{cases} \quad (26)$$

其中, m 表示不同的状态变量的总数。例如本文使用的浅水方程中 m 取 3, 则 Askey 函数表示为:

$$\rho = \begin{pmatrix} \rho_{11} & \rho_{12} & \rho_{13} \\ \rho_{21} & \rho_{22} & \rho_{23} \\ \rho_{31} & \rho_{32} & \rho_{33} \end{pmatrix}, \quad (27)$$

即 ρ 的每个元素 ρ_{ij} 对应于第 i 个变量和第 j 个变量的局地化函数矩阵。由于多元模型引入了一些不同变量之间的联系, 增加了数据同化的复杂度与计算量, 因此多元局地化函数的构造需要考虑这些不同变量之间的联系以及联系的强弱。 u_{ij} 表示第 i 个和第 j 个变量之间的局地化影响参数,

$$u_{ij} = (u_i + u_j) / 2, \quad u_i > 0, \quad i = 1, \dots, m, \quad (28)$$

Askey 函数中的参数 z 和 c 与 GC 函数中定义相同, 分别表示网格点之间的距离和局地化长度尺度。 B 是一个 Beta 函数, s 表示状态变量所在的欧式空间的维度, 而且要满足条件 $v \geq (s + 1) / 2$ 。

2.3 实验设计

本文中假定数值模型完美, 忽略模型误差。真实值轨迹 x^t 可以通过已知的初始条件向前演化得到, 上标 t 表示真实值。观测值 y 通过使用真实值 x^t 和预先给定的观测误差协方差矩阵 R 给出,

$$y = Hx^t + v, \quad v \sim N(0, R) \quad (29)$$

其中, 矩阵 H 是观测算子, 将真实状态变量映射到

算, 时间步为 $\Delta t = 0.01$, 终止时间 T 为 8。

2.2 局地化函数

对于 KS 方程等一元模型, 我们可以使用 GC 函数作为 CL 方法中的局地化函数 ρ , 其表达式如下:

参数。注意, 因子 $\sqrt{10/3}$ 可以调整局地化函数达到最优^[22]。

相反, 多元模型中使用 Askey 函数的一个多元扩展^[21]作为局地化函数 ρ , 形式为,

观测空间。假定观测误差之间不相关, 则矩阵 R 为对角观测误差协方差矩阵, 且对角元素取值为 1。初始集合通过对真实值 x^t 添加 N 次随机误差构造。详细的实验参数设置如下。

KS 模型具有一维周期性, 设定状态向量 u 的维度 n 为 256, 初始真实值可以通过定义在周期域 $0 \leq x \leq 32\pi$ 上的函数 $u = \cos\left(\frac{x}{16}\right) \left[1 + \sin\left(\frac{x}{16}\right)\right]$ 给出。

实验中分别取集合数 N 为 5 或 N 为 10, 观测数 p 是 256 或 p 是 235。数值实验的观测频率(同化间隔)不同, 假定每 5 或 10 个时间步长存在可用的观测值。

对于多元浅水模型, 假定模型定义在矩形网格区域, 每个网格点定义 3 个变量: 水面高度 h , 水平速度 u 和垂直速度 v 。假定水平速度 u 和垂直速度 v 的初始真实状态为零, 即 $u = v = 0$, 水面高度 h 的初始真实状态由如下方式构造,

$$h = \sqrt{x^2 + y^2}, \quad (30)$$

$$h = \sin(h)/h, \quad (31)$$

$$h = \max(h, 0). \quad (32)$$

类似于 KS 模型中的参数设置, 在浅水方程中, 集合数分别取 N 是 5 或 N 是 10, 观测数 p 为 300 或 p 为 240, 每 5 或 10 个时间步长存在可用的观测值。表 1 中给出了上述具体实验参数的配置, 分别在两

个模型中进行 8 组不同的实验, 其中对于 KS 方程全观测 p 为 256, 部分观测 p 为 235; 对于浅水模型全观测 p 为 300, 部分观测 p 是 240。

表 1 不同实验条件的参数

Tab. 1 Parameters for different experimental conditions

实验编号	集合数(N)	观测数(p)	同化间隔
1	5	全观测	5
2	5	部分观测	5
3	5	全观测	10
4	5	部分观测	10
5	10	全观测	5
6	10	部分观测	5
7	10	全观测	10
8	10	部分观测	10

3 实验结果与分析

首先, 分析 GETKF 方法和本文修改的方法(简称 GCL 方法)在一元 KS 模型中的实验结果。为了研究两种方法对预测误差协方差矩阵 P^f 的影响, 以实验 1 为例, 取前两个同化时刻(存在观测数据的时刻), 比较同化过程中两种方法对于预测误差协方差矩阵 P^f 的影响。在图 1 至图 4 的每一幅图中, 从左到右, 从上到下, 依次展示了 GC 局地化函数、预测误差协方差矩阵、局地化函数和预测误差协方差矩阵做舒尔积运算得到的矩阵、使用局地化函数平方根矩阵与预测集合扰动矩阵做舒尔积运算得到的预测误差协方差矩阵。

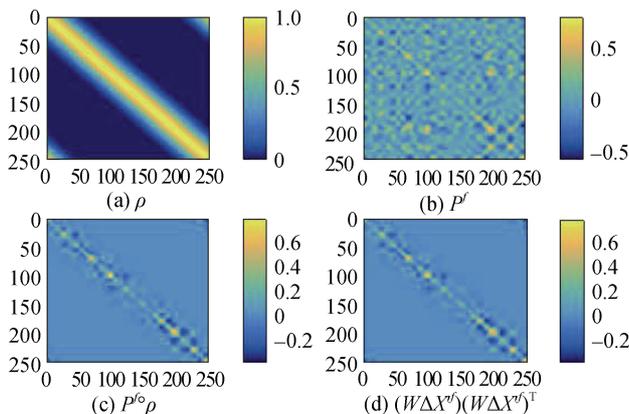


图 1 实验 1 中, 第一个同化时刻 GCL 对预测误差协方差矩阵的影响

Fig. 1 Effect of GCL on the forecast error covariance matrix at the first assimilation time in experiment 1

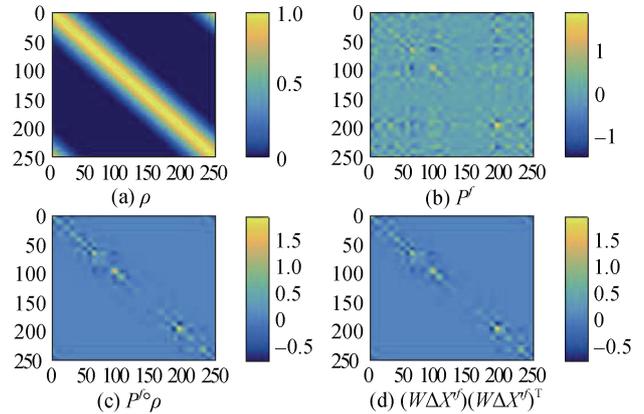


图 2 实验 1 中, 第二个同化时刻 GCL 对预测误差协方差矩阵的影响

Fig. 2 Effect of GCL on the forecast error covariance matrix at the second assimilation time in experiment 1

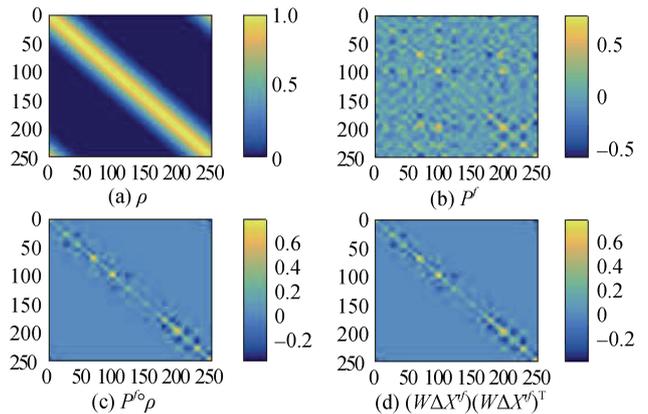


图 3 实验 1 中, 第一个同化时刻 GETKF 对预测误差协方差矩阵的影响

Fig. 3 Effect of GETKF on the forecast error covariance matrix at the first assimilation time in experiment 1

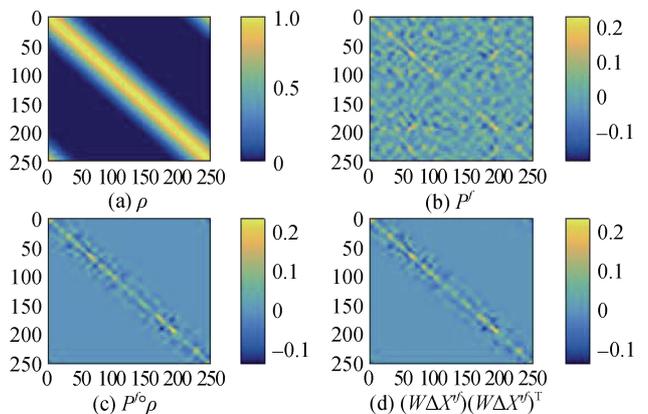


图 4 实验 1 中, 第二个同化时刻 GETKF 对预测误差协方差矩阵的影响

Fig. 4 Effect of GETKF on the forecast error covariance matrix at the second assimilation time in experiment 1

由于 KS 模型具有一维周期性，所以 GC 局地化函数 ρ 是一个对称矩阵，且仅在主对角线附近和矩阵边缘存在较强的相关性。很显然，对于 GETKF 和 GCL 这两种方法，图 1 至图 4 中的第 4 幅子图均显示了使用局地化函数之后协方差矩阵远距离的伪相关得到消除，并且和第 3 幅子图相比效果类似，说明两种方法对于虚假相关性的消除是有效果的。但是，比较图 2 和图 4，可以看出第 2 个时刻两种方法的局地化效果是存在偏差的(图例中数值区间不同)，这可能是源于两种方法的具体实施不同。但是总体上两者对于处理伪相关的效果较好，即截断距离以外的相关性被消除，邻近状态变量和边界上的相关性得到了保留。

以实验 3 和实验 7 为例，分别对 GETKF 方法和 GCL 两种方法的同化效果进行评估。如图 5 和图 6 中不同方法的曲线轨迹所示，大体上可以看出相比于 GETKF 方法的同化效果，本文修改的 GCL 方法更加接近于真实值的轨迹曲线，并且减少了滤波发散的发生，展示了良好的同化效果，说明在一元数值预测模型下，GCL 方法的局地化效果优于 GETKF 方法。

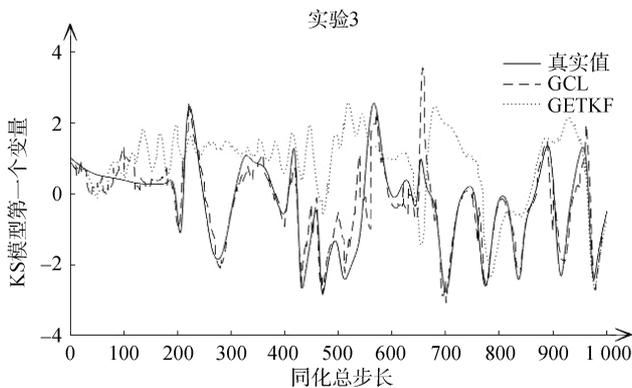


图 5 实验 3, KS 模型第一个变量的同化效果

Fig. 5 Comparison of the assimilation effect of the first variable of the KS model in experiment 3

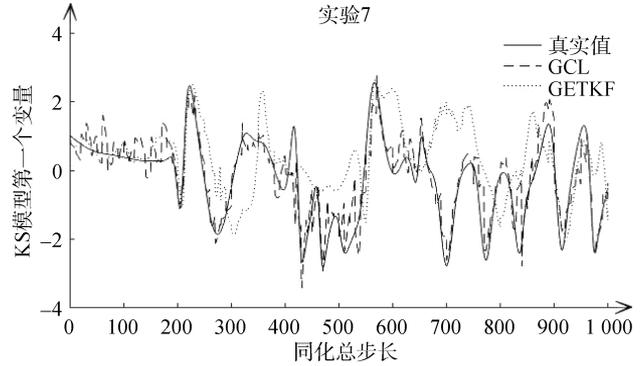


图 6 实验 7 中, KS 模型第一个变量的同化效果

Fig. 6 Comparison of the assimilation effect of the first variable of the KS model in experiment 7

均方根误差(root mean square error, RMSE)可以量化不同局地化方法的效果，直观表示不同方法的优劣程度。模型状态变量的整体均方根误差表示为，

$$RMSE = \sum_{j=1}^n \left(\sqrt{\frac{1}{l} \sum_{i=1}^l (x_{ij}^a - x_{ij}^t)^2} \right) \quad (33)$$

其中, l 表示同化时间的长度, x_{ij}^a 表示第 i 个时刻第 j 个变量的分析值, x_{ij}^t 表示第 i 个时刻第 j 个变量的真实值。此外，由预测集合通过统计意义计算得到预测误差协方差矩阵的秩至多为 $N-1$ ，该矩阵是一个低秩的矩阵，这在传递误差时易造成矩阵的病态，因此需要提升矩阵的秩。理论上 CL 方法对于矩阵秩的提升是有效果的，但是由于 CL 方法应用于 ETKF 方法的局限性，使得效果大打折扣，所以除了计算同化过程中的 RMSE 外，也会对比 GETKF 和 GCL 方法使用前后预测误差协方差矩阵秩的变化。具体如表 2 所示，首先，两种方法的预测误差协方差矩阵的秩与不使用局地化方法相比，显然均有了提高。GCL 方法对于矩阵秩的提高更为明显。其次，GCL 方法的 RMSE 明显小于 GETKF 方法的 RMSE(实验 6 有偏差)，说明在一元模型中，GCL 方法的同化效果要优于 GETKF 方法的同化效果。

表 2 KS 模型中不同实验条件下 RMSE 和矩阵秩的对比

Tab. 2 Comparison of the RMSE and matrix rank under different experimental conditions in the KS model

编号	RMSE(无 CL)	RMSE(GETKF)	RMSE(GCL)	秩(无 CL)	秩(GETKF)	秩(GCL)
1	424.35	308.53	93.21	4	40	100
2	422.03	366.54	112.53	4	40	60
3	411.02	366.13	116.78	4	40	100
4	409.91	350.85	133.00	4	40	100
5	406.93	363.58	116.37	9	90	184

续表

编号	RMSE(无 CL)	RMSE(GETKF)	RMSE(GCL)	秩(无 CL)	秩(GETKF)	秩(GCL)
6	387.79	35.59	126.93	9	90	176
7	412.54	336.49	148.92	9	90	182
8	387.97	320.72	157.46	9	90	189

表 3 中展示了主流的数据同化方法(EnKF)的 RMSE, 采用实验 1 中的观测数和同化间隔, 同时使用变化的集合数进行数值实验。EnKF 方法在集合数 N 是 5 的情况下, RMSE 的数值远远大于 GCL 方法(实验 1), 随着集合数 N 增大到 100, RMSE 的数值在降低, 与 GCL 方法(实验 1)的 RMSE 相当, 但是程序的运行时间也在不断增加。当集合数为 N 取 200 时, EnKF 方法的 RMSE 数值低于 GCL 方法(实验 1), 表现出了良好的数据同化效果, 但是程序运行时间是 GCL 方法使用集合数 N 为 5 时的数倍(实验 1 条件下, GCL 方法运行时间为 4.83 s)。EnKF 通过使用大集合数, 牺牲程序计算时间来获取低 RMSE, 并不是一种很好的处理方式, 并且在真实的数据同化方法应用中, 选取超过 100 个集合是不合适的。上述实验结果表明两种数据同化方法各有优劣, 但是可以看出在集合数较小的情况下, 选择 GCL 方法进行数据同化的效果较好。

表 3 KS 模型中 EnKF 方法的 RMSE
Tab. 3 RMSE of the EnKF method in the KS model

集合数(N)	RMSE	运行时间/s
5	417.23	6.01
10	429.88	8.15
30	399.54	15.79
50	380.61	23.76
70	346.17	31.93
100	165.90	44.97
200	56.33	85.30

最后, 我们将本文提出的 GCL 方法中的 GC 函数替换为 Askey 函数, 扩展方法的适用范围, 在多元浅水模型中检验该方法。表 4 展示了在不同实验条件下, GCL 方法与未使用局地化方法的 RMSE 以及预测误差协方差矩阵的秩的对比。类似于一元数值模型中的实验结果, 使用 GCL 方法后的预测误差协方差矩阵的秩与未使用局地化方法相比, 有了明显的提升, 保证了误差矩阵传递的可靠性, 从而 GCL 方法的 RMSE 与未使用局地化方法相比也有了明显的降低, 说明利用 Askey 函数将 GCL 方法扩展到多

元模型是成功的, 克服了 GETKF 方法只能用于一元模型的局限性。

表 4 浅水模型中不同实验条件下 RMSE 和矩阵秩的对比
Tab. 4 Comparison of the RMSE and matrix rank under different experimental conditions in the shallow water model

编号	RMSE(无 CL)	RMSE(GCL)	秩(无 CL)	秩(GCL)
1	151.29	132.18	4	60
2	151.73	129.33	4	60
3	152.30	128.62	4	60
4	151.91	130.86	4	120
5	108.13	78.81	9	270
6	108.12	87.80	9	270
7	108.88	76.69	9	225
8	108.36	77.55	9	180

4 结论

为解决基于集合的卡尔曼滤波中集合数目过少导致的欠采样问题, 以及 CL 方法对于 ETKF 等平方根卡尔曼滤波方法的不适用, 本文对 GETKF 方法进行了改进, 提出 GCL 方法, 并与 GETKF 方法在相同实验条件下进行了实验对比。结果表明, GCL 方法改善了在一元模型情况中的同化效果, 明显提高了预测误差协方差矩阵的秩, 降低了分析值的均方根误差。此外, 利用一个 Askey 函数实现局地化方法在多元模型中的扩展, 使得提出的 GCL 方法可以用于多元模型中, 克服当前多元模型中局地化方法的欠缺。总体来说, 虽然 GCL 方法在处理欠采样问题以及对于多元模型局地化具有一定的效果, 但是 GCL 方法应用到多元模型时, 为设计多元变量之间的相互作用关系及相关性, Askey 函数需要预先指定的参数较多。在下一步的研究中应找到合适的参数确定依据, 使得提出的新方法具有更广泛的适用性。

参考文献:

[1] EVENSEN G. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics[J]. Journal of Geophysical Re-

- search: *Oceans*, 1994, 99(C5): 10143-10162.
- [2] ANDERSON J L, ANDERSON S L. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts[J]. *Monthly Weather Review*, 1999, 127(12): 2741-2758.
- [3] HOUTEKAMER P, MITCHELL H. A sequential ensemble Kalman filter for atmospheric data assimilation[J]. *Monthly Weather Review*, 2001, 129(1): 123-137.
- [4] WHITAKER J, HAMILL T. Ensemble data assimilation without perturbed observations[J]. *Monthly Weather Review*, 2002, 130(7): 1913-1924.
- [5] HUNT B R, KOSTELICH E J, SZUNYOGH I. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter[J]. *Physica D-Nonlinear Phenomena*, 2007, 230(1-2): 112-126.
- [6] OTT E, HUNT B, SZUNYOGH I, et al. A local ensemble Kalman filter for atmospheric data assimilation[J]. *Tellus Series A-Dynamic Meteorology and Oceanography*, 2004, 56(5): 415-428.
- [7] SCHUR J. Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen[J]. *Journal Für Die Reine Und Angewandte Mathematik*, 1911, 1911(140): 1-28.
- [8] OKE P R, SAKOV P, CORNEY S P. Impacts of localisation in the EnKF and EnOI: experiments with a small model[J]. *Ocean Dynamics*, 2007, 57(1): 32-45.
- [9] JANJIC T, NERGER L, ALBERTELLA A, et al. On domain localization in Ensemble-Based Kalman filter algorithms[J]. *Monthly Weather Review*, 2011, 139(7): 2046-2060.
- [10] GREYBUSH S J, KALNAY E, MIYOSHI T, et al. Balance and ensemble kalman filter localization techniques[J]. *Monthly Weather Review*, 2011, 139(2): 511-522.
- [11] SAKOV P, BERTINO L. Relation between two common localisation methods for the EnKF[J]. *Computational Geosciences*, 2011, 15(2): 225-237.
- [12] MIYOSHI T, YAMANE S. Local ensemble transform Kalman filtering with an AGCM at a T159/L48 resolution[J]. *Monthly Weather Review*, 2007, 135(11): 3841-3861.
- [13] BISHOP C, ETHERTON B, MAJUMDAR S. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects[J]. *Monthly Weather Review*, 2001, 129(3): 420-436.
- [14] PETRIE R E. Localization in the ensemble Kalman filter[D]. Reading: University of Reading, 2008.
- [15] 韩培, 舒红, 许剑辉, 等. 局地化方法在集合转换卡尔曼滤波同化的适用性研究[J]. *地球信息科学学报*, 2016, 18(9): 1184-1190.
HAN Pei, SHU Hong, XU Jianhui, et al. An applicability study of covariance localization method in ETKF data assimilation[J]. *Journal of Geo-information Science*, 2016, 18(9): 1184-1190.
- [16] BISHOP C H, HODYSS D. Ensemble covariances adaptively localized with ECO-RAP. Part 1: tests on simple error models[J]. *Tellus Series A: Dynamic Meteorology and Oceanography*, 2009, 61(1): 84-96.
- [17] BISHOP C H, HODYSS D. Ensemble covariances adaptively localized with ECO-RAP. Part 2: a strategy for the atmosphere[J]. *Tellus Series A: Dynamic Meteorology and Oceanography*, 2009, 61(1): 97-111.
- [18] BISHOP C H, WHITAKER J S, LEI L. Gain form of the ensemble transform Kalman filter and its relevance to satellite data assimilation with model space ensemble covariance localization[J]. *Monthly Weather Review*, 2017, 145(11): 4575-4592.
- [19] GASPARI G, COHN S E. Construction of correlation functions in two and three dimensions[J]. *Quarterly Journal of the Royal Meteorological Society*, 1999, 125(554): 723-757.
- [20] CAMPBELL W F, BISHOP C H, HODYSS D. Vertical covariance localization for satellite radiances in ensemble Kalman filters[J]. *Monthly Weather Review*, 2010, 138(1): 282-290.
- [21] ROH S, JUN M, SZUNYOGH I, et al. Multivariate localization methods for ensemble Kalman filtering[J]. *Nonlinear Processes in Geophysics*, 2015, 22(6): 723-735.
- [22] LORENC A. The potential of the ensemble Kalman filter for NWP-a comparison with 4D-Var[J]. *Quarterly Journal of the Royal Meteorological Society*, 2003, 129(595): 3183-3203.

Covariance localization in the ensemble transform Kalman filter data assimilation method

WANG Ji-chao¹, WANG Yue¹, ZANG Shao-dong¹, YANG Jun-gang², JI Yan-ju¹, RUAN Zong-li¹

(1. College of Science, China University of Petroleum (East China), Qingdao 266580, China; 2. First Institute of Oceanography, Ministry of Natural Resources, Qingdao 266061, China)

Received: Mar. 21, 2020

Key words: covariance localization; ensemble transform Kalman filter; multivariate localization function

Abstract: In ensemble data assimilation, the use of covariance localization (CL) methods is limited. The ensemble transform Kalman filter (ETKF), a variant of ensemble square root filters, is a widely used and computationally efficient data assimilation method. This article theoretically analyzes the difficulties of applying CL to the ETKF method and improves CL to make it more applicable to the ETKF method. Moreover, combined with a shallow water equation, a CL method suitable for the multivariate numerical model is proposed using the Askey function as the multivariate local function. Through specific experimental verification, good analysis results are obtained.

(本文编辑: 丛培秀)