DOI:10.11737/j.issn.1003-0239.2020.05.005

局地误差子空间变换卡尔曼滤波方法的最优参数选取

王昊运[1,2]，王辉[1,3]，张宇[1,3]，万莉颖[1,3]

（1. 国家海洋环境预报中心，北京 100081；2. 中国海洋大学海洋环境与生态教育部重点实验室，山东 青岛 266100；3. 国家海洋局海洋灾害预报技术研究重点实验室，北京 100081）

摘 要： 基于PDAF同化框架，通过Lorenz96模型的孪生实验，探讨了LESTKF同化方案中的两个重要参数局地化半径和遗忘因子对于同化结果的影响。实验结果表明：局地化半径对分析结果的空间分布影响明显。局地化半径过大，不能很好地滤去背景误差协方差矩阵中的虚假相关；局地化半径过小则分析太细节化使得物理量场不符合实际。遗忘因子的选取对于同化效果影响显著：孪生实验证明遗忘因子（取值为0～1）选取适当能够明显提高同化效果，但如果选得太小则会使同化结果过于接近模式，观测信息对模式的调整将减弱。当二者同时作为自变量影响Lorenz96模式的同化效果时，存在一个最优的参数选择区域，但该最优区域紧邻滤波发散的区域，因此在实际同化应用中应格外重视。

关键词： Lorenz96模型；LESTKF同化方法；PDAF同化框架

中图分类号：P456.7 **文献标识码：**A **文章编号：**1003-0239（2020）05-0042-10

1 引言

在大气和海洋模式中，初始状态对于数值模拟的效果至关重要。作为一种顺序同化方法，集合卡尔曼滤波（Ensemble Kalman Filter，EnKF）及其演变的集合同化方法，集诸多优点于一身，是现有同化方法中最具应用发展前景的一个[1-3]。EnKF通过将多个扰动的初始样本作为一个集合[4]，利用这个集合估计背景误差协方差，对卡尔曼滤波方法进行了有效简化。EnKF的优势还在于：其计算代价比卡尔曼滤波和扩展卡尔曼滤波小得多；不要求背景误差协方差是线性演变的；不要求发展模式的线性和伴随模式；可以给集合预报提供好的初始扰动[4-5]。

为了解决EnKF在大气和海洋数值模拟的应用问题，最重要的两个改进就是背景误差协方差的膨胀和局地化分析[6]。通常，模式的状态向量维数很高（为10^7），远远大于集合样本的维数，这会导致背景误差协方差矩阵中的虚假相关。同时因为集合样本离散度的问题，也会导致同化分析对背景误差协方差的低估[7-8]。EnKF及其演变而来的局地化分析在大气和海洋数值模拟中得到了广泛的应用[9-12]。

集合同化的优势在于：集合同化方法不仅给出模式状态的最优估计，而且不需要建立预报模式的切线性和伴随，背景误差协方差“流依赖”。在EnKF同化方法获得成功的基础上，为了不对观测进行扰动，发展了一系列演变、改进的同化方法。例如：集合调整卡尔曼滤波（Ensemble Adjustment Kalman Filter，EAKF）、集合变换卡尔曼滤波（Ensemble Transform Kalman Filter，ETKF）、奇异演变插值卡尔曼滤波（Singular Evolutive Interpolated Kalman Filter，SEIK）、集合平方根滤波（Ensemble Square Root Filter，EnSRF）等。由于顺序同化公式简明、应用相对容易，EnKF及其演变版本得到了快速的发展和应用[13-14]。观测上不加扰动进行同化，能够解

收稿日期：2019-07-02；**修回日期：**2019-09-20。
基金项目：国家重点研发计划（2016YFC1401409）。
作者简介：王昊运（1991-），男，博士在读，主要从事海洋环流数值模拟和资料同化研究。E-mail：wanghaoyun0831@163.com
通讯作者：王辉（1962-），男，研究员，博士，主要从事海洋环流数值模拟、业务化海洋学理论与应用、海洋预报理论与方法研究等。E-mail：wangh@nmefc.cn

决EnKF计算量大和集合成员少时收敛速度慢的问题,这一系列方法被统称为集合均方根滤波[15-17]。例如比EnKF更加高效的SEIK方法,由于采用了二阶取样法,因此可以用更少的集合样本数达到比EnKF更好的效果,同时避免了EnKF对观测向量扰动带来的人为误差,所以更加节省计算资源并且更加有效。实际应用中决定集合同化方法计算代价最关键的就是集合样本数,它直接决定了该同化方法需要在预报步时积分模式的次数。Nerger等[18]结合了ETKF在低维集合样本展开的子空间对误差协方差矩阵估计的优势和SEIK采用二阶取样法可以带来的小样本数的优势,提出了误差子空间变换卡尔曼滤波(Error Subspace Transform Kalman Filter,ESTKF)。ESTKF被证明是一种更加高效的误差子空间滤波,其局地化分析方案称为局地误差子空间变换卡尔曼滤波(Local Error Subspace Transform Kalman Filter,LESTKF)[19]。

LESTKF在实际应用中,同样需要解决背景误差协方差矩阵的低估和因为计算能力不足导致集合样本数过小引起的虚假相关。因此局地化方案中的重要参数"局地化半径"和LESTKF分析方案中用来膨胀背景误差协方差矩阵的"遗忘因子"(Forgetting Factor)这两个重要参数对同化效果起着决定性的作用[19]。研究这两个参数如何影响LESTKF的同化性能和同化效果,以及如何选取这两个参数才能最小化分析误差,对于将该同化方法应用于实际的大气海洋模式是十分关键的[20]。本文利用Lorenz96模型结合LESTKF同化方法,通过"局地化半径"和"遗忘因子"设置孪生实验,研究这两个重要参数对于同化效果的影响。

本文将首先探究"遗忘因子"对背景误差协方差膨胀效果及同化结果的影响;其次将研究"局地化半径"和"遗忘因子"的共同影响;最后结合Lorenz96模型孪生实验的结果分析这两个重要参数在集合同化分析中扮演的角色,并针对如何优化选取参数提出建议和总结。

2　同化方法及同化框架

2.1　ESTKF同化方法介绍

对于非线性海洋大气系统而言,海洋大气在t_k时刻的n维状态向量为x_k,和它对应的误差协方差矩阵为$\boldsymbol{P}_K$。m个集合成员组成的状态向量集合可表示为$x_k^{(\alpha)},\alpha=1,\cdots,m$。用集合均值来表示对$t_k$时刻的状态估计:

$$\overline{\boldsymbol{x}_k}=\frac{1}{m}\sum_i^m \boldsymbol{x}_k^{(i)} \tag{1}$$

集合样本矩阵为:

$$\boldsymbol{X}_k=\left\lfloor \boldsymbol{x}_k^{(1)},\cdots,\boldsymbol{x}_k^{(m)}\right\rfloor \tag{2}$$

背景误差协方差矩阵:

$$\boldsymbol{P}_k=\frac{1}{m-1}\boldsymbol{X}_k'(\boldsymbol{X}_k')^{\mathrm{T}} \tag{3}$$

集合扰动由下式给出:

$$\boldsymbol{X}_k'=\boldsymbol{X}_k-\overline{\boldsymbol{X}_k} \tag{4}$$

$$\overline{\boldsymbol{X}_k}=\left\lfloor \overline{\boldsymbol{x}_k},\cdots,\overline{\boldsymbol{x}_k}\right\rfloor \tag{5}$$

同化预报时首先将集合样本分别通过模式积分至同化时刻。

观测向量记为$\boldsymbol{y}_k^o=\boldsymbol{H}_k(\boldsymbol{x}_k^f)+\boldsymbol{\varepsilon}_k$,为$p$维向量,线性观测算子$\boldsymbol{H}$将背景场投影在观测空间,观测误差记为$\boldsymbol{\varepsilon}_k$。在集合同化方法中,认为观测误差服从高斯分布,观测误差协方差矩阵记为$\boldsymbol{R}$。

在ESTKF分析步中,背景误差协方差$\boldsymbol{P}^f$在形式上用集合样本$\boldsymbol{X}^f$表示,即:

$$\boldsymbol{P}^f=\boldsymbol{LGL}^{\mathrm{T}} \tag{6}$$

$$\boldsymbol{L}=\boldsymbol{X}^f\tilde{\boldsymbol{T}} \tag{7}$$

$$\boldsymbol{G}=(m-1)^{-1}(\tilde{\boldsymbol{T}}^{\mathrm{T}}\tilde{\boldsymbol{T}})^{-1} \tag{8}$$

式(7)和(8)中:$\tilde{\boldsymbol{T}}$是一个$m\times(m-1)$维的满秩矩阵,并且每列元素的和为零,即

$$\tilde{\boldsymbol{T}}=\begin{bmatrix}\boldsymbol{I}_{(m-1)\times(m-1)}\\ \boldsymbol{0}_{1\times(m-1)}\end{bmatrix}-\frac{1}{m}\left[\boldsymbol{I}_{(m-1)\times(m-1)}\right] \tag{9}$$

$\tilde{\boldsymbol{T}}$矩阵的作用是在计算矩阵$\boldsymbol{L}$时剔除集合样本矩阵X^f的集合平均,即计算集合扰动。值得注意的是,矩阵$\boldsymbol{L}$是个$n\times(m-1)$维矩阵,只存储前$m-1$个集合扰动。

分析场通过集合扰动矩阵$\boldsymbol{L}$给出:

$$\overline{\boldsymbol{x}^a}=\overline{\boldsymbol{x}^f}+\boldsymbol{L}\bar{\boldsymbol{w}} \tag{10}$$

$$\bar{\boldsymbol{w}}=\tilde{\boldsymbol{A}}(\boldsymbol{HL})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{y}^0-\boldsymbol{H}\overline{\boldsymbol{x}^f}) \tag{11}$$

$$\tilde{\boldsymbol{A}}^{-1}=\tilde{\rho}\boldsymbol{G}^{-1}+(\boldsymbol{HL})^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{HL} \tag{12}$$

式(10)—(12)中:权重向量$\bar{\boldsymbol{w}}$、变换矩阵$\tilde{\boldsymbol{A}}$分别为$(m-1)$维、$(m-1)\times(m-1)$维。参数取值范围为$0<\tilde{\rho}\leqslant 1$,称作"遗忘因子",用来放大背景误差协方

差。$\tilde{\boldsymbol{\rho}}$作为该同化方法的重要参数，其取值直接影响着ESTKF的同化效果，也是本文中的主要研究对象之一。

该方法在得到分析场后需要进行再取样。由于集合样本矩阵经过了变换用$\tilde{\boldsymbol{X}}^a$表示，$\tilde{\boldsymbol{P}}^a$同理。

$$\tilde{\boldsymbol{X}}^a = \overline{\tilde{\boldsymbol{X}}^a} + \sqrt{m-1}\,\boldsymbol{L}\tilde{\boldsymbol{C}}\boldsymbol{\Omega}^{\mathrm{T}} \tag{13}$$

根据之前的研究，在SEIK中，矩阵$\tilde{\boldsymbol{C}}^{-1}$由$\tilde{\boldsymbol{A}}^{-1}$经过Cholesky分解得到，即$(\tilde{\boldsymbol{C}}^{-1})^{\mathrm{T}}\tilde{\boldsymbol{C}}^{-1} = \tilde{\boldsymbol{A}}^{-1}$。$\boldsymbol{\Omega}$是一个$m \times (m-1)$维的矩阵，该矩阵所有列向量相互正交，并且与$(1,\cdots,1)^{\mathrm{T}}$也正交。矩阵$\boldsymbol{\Omega}$的作用是与$(m-1)\times(m-1)$维的集合变换矩阵$\tilde{\boldsymbol{A}}^{-1}$重新生成新的集合扰动矩阵。在ESTKF中，$\boldsymbol{\Omega}$矩阵记为$\hat{\boldsymbol{\Omega}}$，即：

$$\hat{\boldsymbol{\Omega}}_{i,j} = \begin{cases} 1 - \dfrac{1}{m}\dfrac{1}{\dfrac{1}{\sqrt{m}}+1}, i = j, i < m \\ -\dfrac{1}{m}\dfrac{1}{\dfrac{1}{\sqrt{m}}+1}, i \neq j, i < m \\ -\dfrac{1}{\sqrt{m}}, i = m \end{cases} \tag{14}$$

式中：$\hat{\boldsymbol{\Omega}}$为Householder矩阵[18]。$\hat{\boldsymbol{\Omega}}$的作用为将$\boldsymbol{X}^f$（$n\times m$维）集合样本空间中的向量投影在矩阵$\boldsymbol{L}$（m−1个集合扰动）所在的误差子空间。同样矩阵$\hat{\boldsymbol{\Omega}}$是满秩矩阵，并且列向量和为零。将式（8）—（13）中的$\tilde{\boldsymbol{T}}$矩阵替换为$\hat{\boldsymbol{\Omega}}$矩阵，最终算法就是ESTKF。

除了传统的EnKF等滤波的优势之外，ESTKF的本质是在SEIK的基础上将集合变换矩阵的计算通过$\hat{\boldsymbol{\Omega}}$矩阵投影在误差子空间来完成。该方法相对于SEIK来说：计算矩阵$\boldsymbol{L}$时，不必像之前一样忽略最后一列的集合样本，同时剔除了集合成员顺序的影响；ESTKF由于在误差子空间计算，所以计算代价要小于SEIK。

2.2 局地化方案

为了消除集合样本数不足导致的背景误差协方差矩阵中的虚假相关，假设只有距离模式格点在一定范围（即局地化半径）内的观测才会对模式格点产生影响。局地化不仅可以减小虚假相关对同化效果的影响，还可以减小计算量，尤其当观测数量远远大于模式集合样本数时。而且局地化可以保证集合同化方法在高维模式应用上的合理性，即对于每个模式格点而言，模式状态的调整是在一个相对较大的集合空间内实现的。需要注意的是，由于局地化半径是为了消除虚假相关，而不同模式可能有不同的虚假相关，所以局地化半径随模式不同而不同。同时，局地化可能会在局地区域的边缘引起模式状态的不连续。为了解决这一问题，通常引入平滑的、以局地化半径为参数、以观测和格点距离为自变量的局地化函数，使得观测的影响随距离逐渐衰减。本研究中选用的局地化函数是五阶Gaspari-Cohn相关函数[17]，即：

$$\begin{cases} -\dfrac{1}{4}\left(\dfrac{d}{r_{\mathrm{loc}}}\right)^5 + \dfrac{1}{2}\left(\dfrac{d}{r_{\mathrm{loc}}}\right)^4 + \dfrac{5}{8}\left(\dfrac{d}{r_{\mathrm{loc}}}\right)^3 - \dfrac{5}{3}\left(\dfrac{d}{r_{\mathrm{loc}}}\right)^2 + 1, \\ 0 \leqslant d < r_{loc} \\ \dfrac{1}{12}\left(\dfrac{d}{r_{\mathrm{loc}}}\right)^5 - \dfrac{1}{2}\left(\dfrac{d}{r_{\mathrm{loc}}}\right)^4 + \dfrac{5}{8}\left(\dfrac{d}{r_{\mathrm{loc}}}\right)^3 + \dfrac{5}{3}\left(\dfrac{d}{r_{\mathrm{loc}}}\right)^2 - \\ 5\left(\dfrac{d}{r_{\mathrm{loc}}}\right) + 4 - \dfrac{2}{3}\left(\dfrac{d}{r_{\mathrm{loc}}}\right)^{-1}, \quad r_{\mathrm{loc}} \leqslant d < 2r_{\mathrm{loc}} \\ 0, d \geqslant 2r_{\mathrm{loc}} \end{cases} \tag{15}$$

式中：d表示观测到模式格点的距离，r_{loc}表示局地化半径，计算结果表示观测对分析点的权重。利用Gaspari-Cohn相关函数按照权重结合局地化半径中的观测得到该格点处的分析值。

参数$\tilde{\boldsymbol{\rho}}$的选取对同化的影响主要发生在LESTKF同化的分析步中，在协方差矩阵的计算中引入了一个膨胀因子来增加滤波稳定性。因为模式的强非线性，导致在相空间中对初值十分敏感，在同化中太过相信模式预报的背景场将容易导致滤波发散，因此该参数的选取对于同化效果有着重要的影响。

局地化半径通过把同化的区域分解成小的子区域，并行地同时更新每个格点的分析，更新时仅用到这个格点某一半径内的所有观测。在分析局地化的基础上，对该半径内的不同观测引入权重的局地化方法被称为观测局地化（Observation Localization，OL）。研究表明，在局地化分析时，针

对不同的集合样本数局地化半径存在一个最优的选择[18]。

2.3　并行数据同化框架(Parallel Data Assimilation Framework,PDAF)

集合同化方法作为一种顺序同化方法,在海洋和大气领域得到广泛应用。EnKF、集合最优插值(Ensemble Optimal Interpolation,EnOI)提出后,为了解决计算代价等问题,集合平方根滤波等分析方法演变出一系列的集合同化方法,如ETKF、EAKF、减秩卡尔曼滤波(Singular Evolutive Extended Kalman Filter,SEEK)、SEIK、ESTKF等。这些集合同化方法都是先由预报步分别积分集合样本;分析步仅需要模式提供的部分信息,通常只依赖于状态向量,而不是单个物理量场。例如在海洋模式的状态向量中,存储着U、V、z、S、T等模式变量场,或是需要估计的模式参数。对于观测算子H的计算,只需要知道观测位置在状态向量中的存放位置即可。以上属性使得建立一个用通用方式实现集合同化方法的核心算法,并通过调用通用接口来进行同化的同化框架成为可能。这将极大地降低同化方案的实施。由于通用接口的设定,同化参数也便于系统化的管理和调整。

PDAF同化框架就是一些集合同化方法的算法库(网址:http://pdaf.awi.de)[21]。目前全球大部分同化系统都是离线进行的,也就是模式集合积分与同化分析步分两个程序进行。这种同化方式虽稳定但低效。因为模式积分程序和同化程序之间需要用文件来传递信息,这种运行方式称为"离线模式"。另一种运行方式为直接将同化程序写进模式代码中整合成为一个程序,这种运行方式更高效,但是需要调整很小一部分模式代码,这种运行方式称为"在线模式"。前者是将模式积分和同化分析步分两个单独程序运行,模式的输出结果输入到同化程序,同化的分析场又作为模式的启动场;后者是将同化程序与模式耦合起来,需要对模式代码做扩展,调用PDAF核心函数,形成一个完整的运行程序。本研究中采用PDAF同化框架结合Lorenz96以"在线模式"运行。实际情况中由于模式复杂,往往采用"离线模式"[21-23]。

3　Lorenz96-PDAF孪生实验

3.1　实验设置及PDAF参数选取

Lorenz96模型作为大气和海洋的低阶近似,具有非线性,存在混沌吸引子。Lorenz96已经被广泛应用于大气海洋预测和资料同化的研究中[24]。其非线性强,同时对初值非常敏感,其动力框架为:

$$\mathrm{d}X_i/\mathrm{d}t = (X_{i+1} - X_{i-2})X_{i-1} - X_i + F \quad (16)$$

式中:X_i为模型变量,F为定常强迫,F的大小决定着非线性的程度。本文的实验设计中,模型维数等于dim_state=40,即i = 1,2,…,40,强迫项F选为8,数值积分采用四阶龙格–库塔格式。

对于孪生实验而言,将模式积分10 000步得到的状态向量的时间序列作为"真值",并用第1 001步时的模式状态作为初始场(前1 000步为模式调整阶段,即"spin up"阶段),其状态向量序列即为40×10 000维的矩阵。

观测序列是在Lorenz96模式"真值"序列的基础上,加上不相关的随机噪声生成的。本节中,为了探索不同同化强度下遗忘因子和局地化半径的影响,扰动了不同的观测序列,两个观测序列的标准差分别取为1.0和0.1。为了去除模式调整阶段的影响,本实验只同化1 000步之后的观测,采用积分一步同化一步观测的方式。

用于集合同化的初始集合样本由二阶取样法(Second-order Exact Sampling)在10 000步模式"历史真值"的基础上生成初始样本。本节为了探究遗忘因子和局地化半径在不同样本数下的同化效果,分别用10个、30个样本做了对照试验。集合孪生同化实验流程见图1。

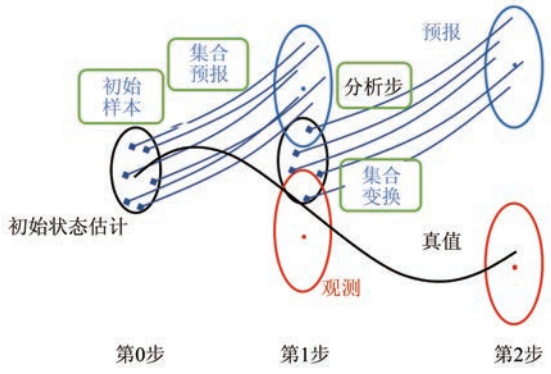


图1　集合孪生同化实验示意图

由于本孪生实验的Lorenz96模型和同化方法均在PDAF同化框架中进行，表1给出了主要同化参数的设置和选项。其中dim_ens为集合样本数，forget和local_range分别对应“遗忘因子”和“局地化半径”。

表1 同化实验主要参数

参数	参数描述	取值
dim_ens	集合样本数	30或10
forget	遗忘因子	0.5～1.0
step_null	实验跳过的时间步，“spin up”步数	1 000
total_steps	同化实验总积分步数	10 000
local_range	局地化半径	0～20个格点
locweight	局地化权重函数	五阶GC函数
rms_obs	观测误差	1.0和0.1

3.2 遗忘因子对同化效果的影响

为了探究不同遗忘因子对同化结果的影响，本实验采用local_range为5个模式格点、rms_obs=1.0、dim_ens=30；但将遗忘因子取不同值。将同化结果、模式预报结果与模式真值、估计值（30个样本的集合平均）做均方根误差（Root Mean Square Error，RMSE）对比，综合探究遗忘因子对同化结果的影响。为了方便表述，以下遗忘因子均记为f值。

图2给出了同化至1 000步，即模式第2 000步的真值、观测值、估计值和分析场。图2a和2b分别为第2 000步的真值和观测值；图2c和2d分别为f=1.0时模式估计场和同化分析场；图2e和2f分别为f=0.9时的模式估计场和同化分析场。

通过对比发现模式第40个格点上的观测值相对真值出现了较大扰动（见图2b、2d、2f的红框部分）。而只有f=0.9的分析场有一个抬升，向观测值做出了调整，这正是因为f对背景误差协方差的膨胀作用导致的分析结果向观测偏移。

为了研究f对分析场和预报场的影响，本实验又将f的取值扩大为0.5～1.0，并计算其同化5 000步后与真值的RMSE。

图3分别为预报场和分析场的RMSE，横坐标为同化步（因为时间步过于密集，因此每隔50步填值，横坐标100即代表第5 000步，以此类推）。f从上到下由0.5（协方差膨胀最大）递增至1.0（协方差不膨胀）。

对于单个RMSE序列来说，在同化开始时

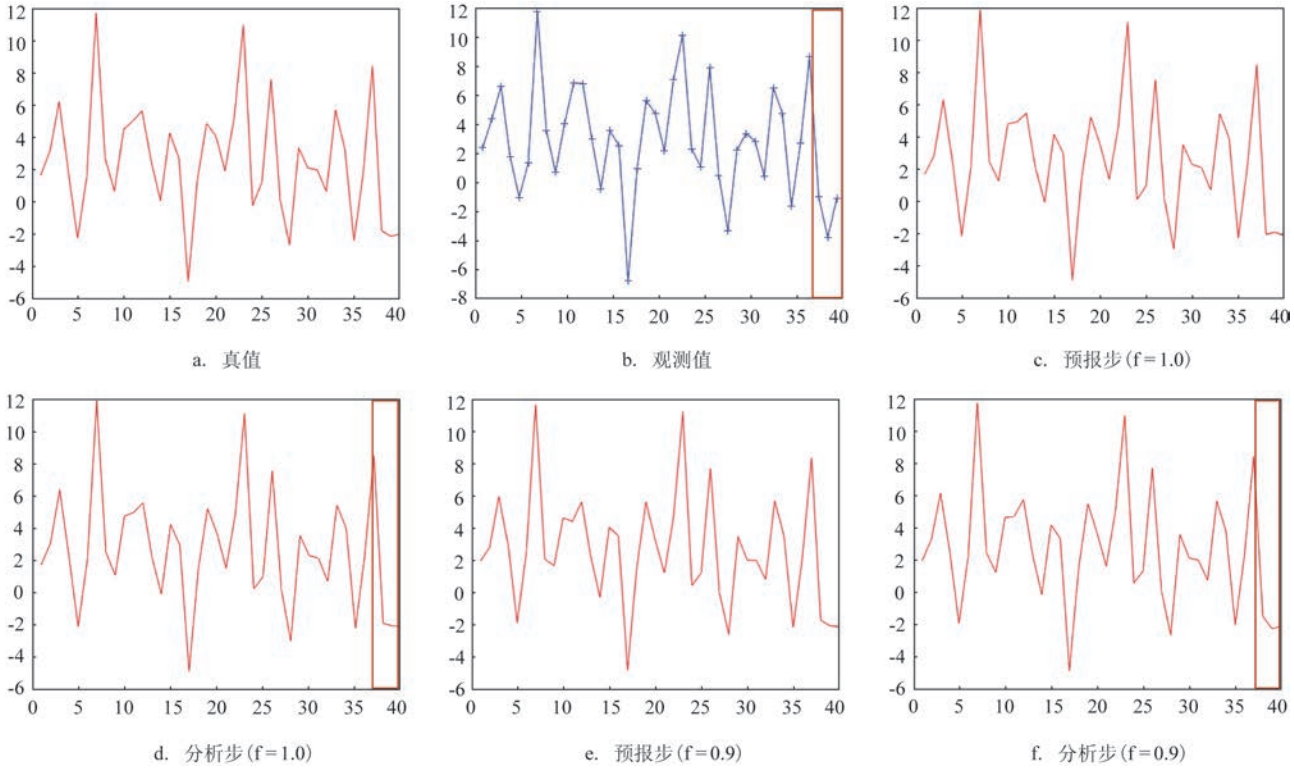


图2 同化实验1 000步结果（横坐标代表模式格点，纵坐标代表对应值）

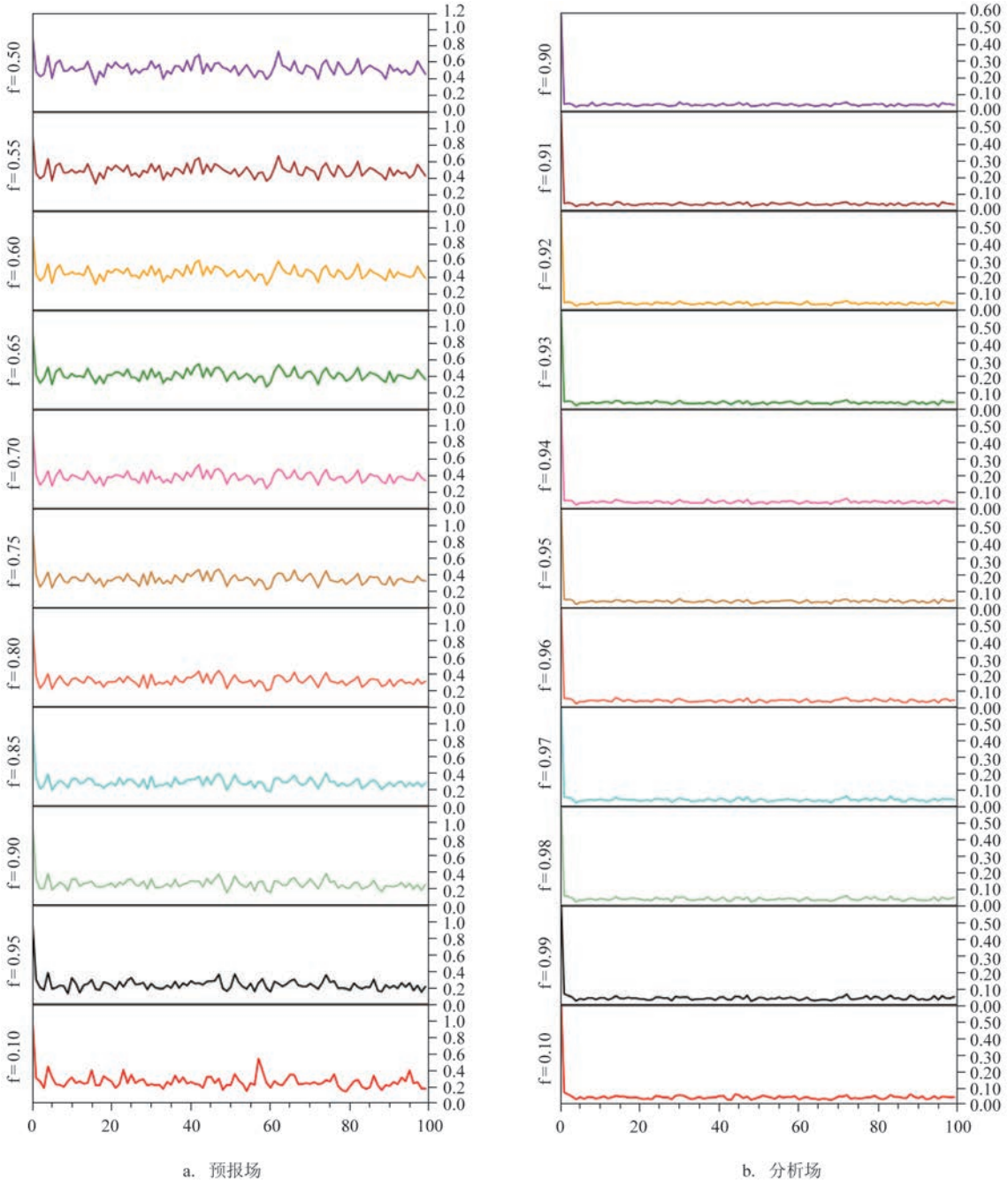


图3　预报场和分析场RMSE(横坐标为缩放50倍的时间步,纵坐标为与真值的RMSE)

RMSE最大。这是因为初始样本是从模式之前的“历史状态”中提取的,与真值的误差最大;但随着同化预报步和分析步的交替,预报场和分析场均迅速向真值收敛。但是随着同化实验的进行会出现误差的突然增长。该特征与Pham[8]的图1一致。这一现象可能由于Lorenz系统进入了强的非线性区域所导致。

通过图3a、b的对比,发现LESTKF对于所有实验而言均对预报场有所改善,但不同f值的改善程度不同。通过纵向比较可以看出:f为0.95与f为1.0时,f的使用明显避免了同化至3 000步左右(图中横坐标约60)的误差爆发;f为0.5时误差最大;从上到下RMSE有一个先减小再增大的过程。

为了探究遗忘因子整体对同化结果的影响,分

别对5 000步同化实验的预报场和分析场的RMSE求时间平均(见图4)。结果表明:f =0.95时分析场和预报场的误差最小;随着f的减小,误差随之增大。f =1.0时分析误差和背景误差均存在着低估,随着f减小低估得到改善;但随着f继续减小,则出现了分析误差和背景误差的高估,导致同化结果过于接近观测从而偏离真值。蓝线和黑线的交点则是最优f值。这说明遗忘因子通过控制对背景误差协方差的估计会显著影响同化效果,高估和低估均不是最优f值。

图4a、b的对比表明对于固定的f而言,分析场相对于真值的RMSE明显小于预报场相对于真值的RMSE。例如当f =0.95时,预报场相对于真值的RMSE约为0.25,而分析场相对于真值的RMSE约为0.2。这说明f通过对背景误差协方差的放大的确可以改善LESTKF的同化效果;但随着f越来越小,分析场对于预报场的提升依然存在,但提升效果开始变小。

3.3 局地化半径与遗忘因子的共同影响

本实验根据Nerger等[18]的实验设置,将total_steps设置为60 000步,得到的结果作为真值;观测依然在真值的基础上增加随机扰动,观测误差分别以1.0和0.1的标准差生成。前1 000步仍然作为“spin up”阶段。本实验将通过不同的观测误差控制同化实验的强度。集合样本分别取10个和30个。因为在实际的集合同化实验中,往往因为计算资源的限制,不会选取过多的集合样本,因此选取10个集合样本更加接近实际情况。

local_range的范围为0～20个格点,这是因为:Lorenz96模式共有40个格点,并且是周期边界条件;local_range超过20个格点后相当于没有局地化分析。将不同f和local_range组合进行同化实验,结果依然用RMSE的时间平均来表示。当时间平均RMSE大于设定的观测误差时,认为发生了滤波发散。

图5a给出了30个集合样本、观测误差为1.0的组合实验结果;5b为10个集合样本、观测误差为1.0的组合实验结果;5c为10个集合样本但观测误差降低至0.1的组合实验结果。格点上的数字代表二者组合实验分析场的时间平均RMSE,白色格点说明该参数组合实验结果的误差超过了设定的扰动的观测误差,发生了滤波发散,因此不填色。

图5可以看出当集合样本数为30个时,不论local_range和f如何选取,均没有发生滤波发散,而且存在一个最优的参数组合区域。这表明当集合样本数足够大、同化方法对背景误差协方差的估计很精确时,引入遗忘因子和局地化分析依然对提高同化结果有帮助,能够找到一个最优的参数搭配区域。例如图5a中local_range＞6、f ＞0.93的区域。

当集合样本数为10个时,这种情形更加接近实际同化,如图5b、c中出现了大面积的滤波发散区域。值得注意的是,最优的参数选择区域和滤波发散区域非常接近,两者紧邻。例如local_range = 7、f = 0.98时,RMSE约为0.2;但是当local_rangen = 8、f仍然为0.98时,则出现了RMSE为3.5的滤波发散,

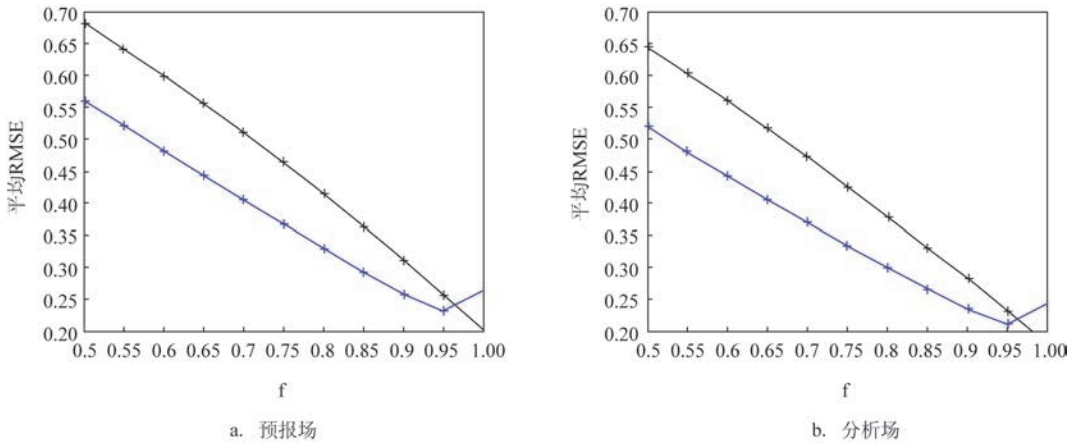
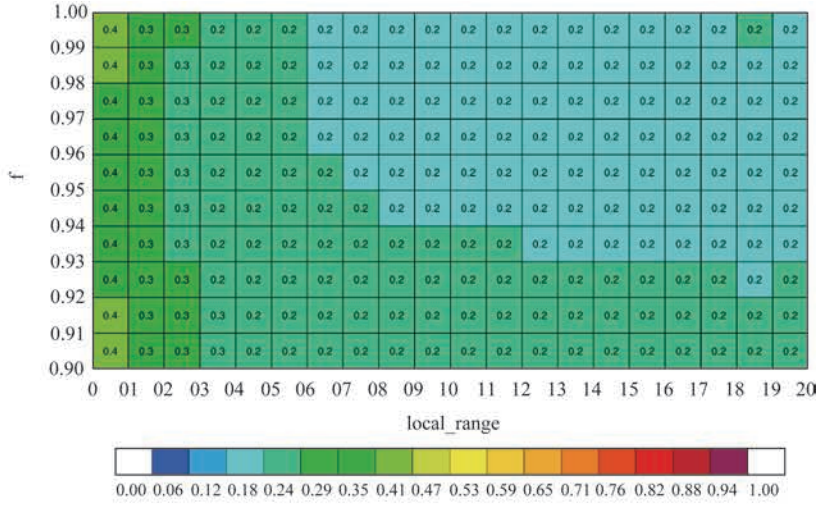
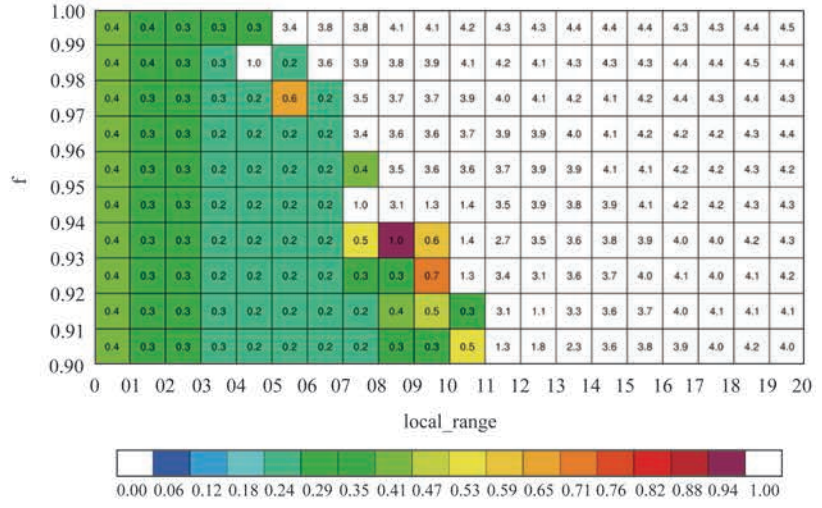


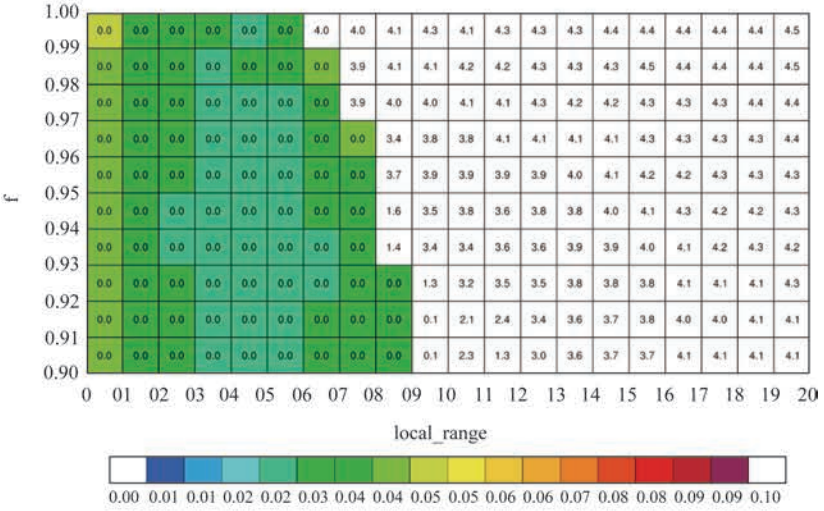
图4 时间平均RMSE(蓝线代表与真值的误差,黑线代表与估计真值的误差)



a. 观测误差为1.0, n=30



b. 观测误差为1.0, n=10



c. 观测误差为0.1, n=10

图5　组合实验结果(表中数字代表RMSE,白色区域为滤波发散区)

这个结果与Nerger等[18]的结果仍然一致。Nerger解释是由初始样本随机选取引起的同化实验前期的过渡阶段导致的,差的随机初始样本会使该过渡阶段变长,因此导致滤波发散。

为了探讨观测误差对于同化实验的影响,将观测重新用0.1的标准差随机生成,并同样用10个集合样本做了组合实验,结果见图5c。当通过减小观测误差来增加同化强度时,滤波发散的区域有所增大,但依然存在一个最优的参数选择区域。

4 总结与讨论

本文通过Lorenz96模型的孪生实验,分别验证了局地化半径和遗忘因子的不同选取对同化分析场的影响,并且详细介绍了PDAF同化中一些参数的设置。结论如下:

(1)局地化半径对分析结果的空间分布影响明显:局地化半径过大,不能很好地滤去背景误差协方差矩阵中的虚假相关;局地化半径过小则分析太细节化,使得物理量场不符合实际。

(2)遗忘因子作为单独影响同化实验的因子时,分析误差随着f值的增大有一个先减小再增大的过程。这是因为f值控制着对分析误差协方差和背景误差协方差的估计,对真值的低估和高估均会引起误差。图4中可以清楚的看到,估计的RMSE与实际的RMSE的交点就是理论上最优的遗忘因子。但在实际同化应用中,由于对真值的估计不准确,因此对于参数f的选取只能通过实验方法去寻找。

(3)f的选取对于同化效果影响显著。Lorenz96理想实验结果表明:f的作用是人为放大背景误差协方差,防止滤波发散。f取值并非越小越好,太小会使同化结果过于接近模式,从而减弱观测信息对模式的调整。f恰当选取(取值为0~1),则可以明显提高同化效果。因此在实际同化中选取该参数时应格外注意。

(4)遗忘因子和局地化半径作为共同因子影响同化时,的确存在一个最优的区域。但是最优区域的选取需要慎重,因为最优的参数组合的区域往往在滤波发散的临界区域附近。一旦选取不当,则很容易出现滤波发散,这在实际同化中一定要注意。本文为今后进一步利用LESTKF实际同化业务应用中参数的最优化选取做了铺垫。

作为仍在不断发展的同化方案,ESTKF及PDAF同化框架已经得到越来越广泛的应用。由于PDAF同化框架接口化、开源的特性,可以更方便地研究分析步中的其他参数并实现优化。例如,如何选取自适应的局地化半径(随纬度、同化的物理量而改变),以及如何进一步选取自适应的遗忘因子,以便协方差膨胀随着同化问题不同而变化,是值得进一步研究的问题。

资料同化是一个综合问题,不仅需要同化方案的演进,同时还需要计算技术的进步,二者缺一不可。怎样在不提高计算代价的前提下提高同化效果,是资料同化理论研究中的一个重要方向。是否能够提出一个更有效率的同化方案,并且能拥有集合滤波中随时间演变的背景误差协方差矩阵,仍是资料同化研究的一个重要问题。

参考文献:

[1] Evensen G. Advanced data assimilation for strongly nonlinear dynamics[J]. Monthly Weather Review, 1997, 125(6): 1342-1354.

[2] Evensen G, Van Leeuwen P J. An ensemble Kalman smoother for nonlinear dynamics[J]. Monthly Weather Review, 2000, 128(6): 1852-1867.

[3] Evensen G. The ensemble Kalman filter: theoretical formulation and practical implementation[J]. Ocean Dynamics, 2003, 53(4): 343-367.

[4] Evensen G. Data assimilation: the ensemble Kalman filter[M]. Berlin, Heidelberg: Springer, 2007.

[5] Evensen G. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics[J]. Journal of Geophysical Research: Oceans, 1994, 99(C5): 10143-10162.

[6] Houtekamer P L, Mitchell H L. Data assimilation using an ensemble Kalman filter technique[J]. Monthly Weather Review, 1998, 126 (3): 796-811.

[7] Pham D T, Verron J, Gourdeau L. Singular evolutive Kalman filters for data assimilation in oceanography[J]. Comptes Rendus de l'Academie des Sciences Series II A Earth and Planetary Science, 1998, 326(4): 255-260.

[8] Pham D T. Stochastic methods for sequential data assimilation in strongly nonlinear systems[J]. Monthly Weather Review, 2001, 129 (5): 1194-1207.

[9] Nerger L, Danilov S, Hille W, et al. Using sea-level data to constrain a finite-element primitive-equation ocean model with a local SEIK filter[J]. Ocean Dynamics, 2006, 56(5-6): 634-649.

[10] Losa S N, Danilov S, Schröter J, et al. Assimilating NOAA SST

data into BSH operational circulation model for the North and Baltic Seas: Part 2. Sensitivity of the forecast's skill to the prior model error statistics[J]. Journal of Marine Systems, 2014, 129: 259-270.

[11] Bell M J, Lefèbvre M, Le Traon P Y, et al. GODAE: The global ocean data assimilation experiment[J]. Oceanography, 2009, 22 (3): 14-21.

[12] Whitaker J S, Hamill T M. Ensemble data assimilation without perturbed observations[J]. Monthly Weather Review, 2002, 130 (7): 1913-1924.

[13] Bishop C H, Etherton B J, Majumdar S J. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects [J]. Monthly Weather Review, 2001, 129(3): 420-436.

[14] Courtier P, Thépaut J N, Hollingsworth A. A strategy for operational implementation of 4D-Var, using an incremental approach [J]. Quarterly Journal of the Royal Meteorological Society, 1994, 120(519): 1367-1387.

[15] Hamill T M, Whitaker J S, Snyder C. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter[J]. Monthly Weather Review, 2001, 129(11): 2776-2790.

[16] Houtekamer P L, Mitchell H L. A sequential ensemble Kalman filter for atmospheric data assimilation[J]. Monthly Weather Review, 2001, 129(1): 123-137.

[17] 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.

[18] Nerger L, Janjić T, Schröter J, et al. A regulated localization scheme for ensemble-based Kalman filters[J]. Quarterly Journal of the Royal Meteorological Society, 2012, 138(664): 802-812.

[19] Nerger L, Hiller W, Schröter J. A comparison of error subspace Kalman filters[J]. Tellus A, 2005, 57(5): 715-735.

[20] Nerger L, Gregg W W. Assimilation of SeaWiFS data into a global ocean-biogeochemical model using a local SEIK filter[J]. Journal of Marine Systems, 2007, 68(1-2): 237-254.

[21] Nerger L, Hiller W, Schröter J. PDAF - the parallel data assimilation framework: experiences with Kalman filtering[M]//Zwieflhofer W, Mozdzynski G. Use of High Performance Computing in Meteorology. Reading: World Scientific, 2005: 63-83.

[22] Nerger L, Hiller W. Software for ensemble-based data assimilation systems-Implementation strategies and scalability[J]. Computers & Geosciences, 2013, 55: 110-118.

[23] Sanchez S, Wicht J, Bärenzung J, et al. Sequential assimilation of geomagnetic observations: perspectives for the reconstruction and prediction of core dynamics[J]. Geophysical Journal International, 2019, 217(2): 1434-1450.

[24] Lorenz E N. Predictability-a problem partly solved[C]//Proceedings Seminar on Predictability. Reading, UK: ECMWF, 1996: 1-18.

The choice of the optimal parameters in a Local Error Subspace Transform Kalman Filter

WANG Hao-yun[1,2], WANG Hui[1,3], ZHANG Yu[1,3], WAN Li-ying[1,3]

(*1. National Marine Environmental Forecasting Center, Beijing 100081 China; 2. Laboratory of Marine Environment and Ecology, Ocean University of China, Qingdao 266100 China; 3. Key Laboratory of Research on Marine Hazards Forecasting, National Marine Environmental Forecasting Center, Beijing 100081 China*)

Abstract: Based on the ensemble assimilation method LESTKF within PDAF, this paper analyzes the impact of local radius and forgetting factor on assimilation results using twin experiments with Lorenz96 model. The results show that the localization radius significantly impacts the spatial distribution of the analysis results. The spurious correlations in the background error covariance can't be well filtered if the localization radius is too large, while the physical quantities field doesn't conform with the reality due to over-detailed analysis if the localization radius is too small. The twin experiments reveal that the assimilation performance could be significantly improved by choosing proper forget factor (0～1). The smaller the forget factor is, the closer the assimilation results will resemble the model simulation, which indicates the weakening impacts of observations in adjusting the model. Therefore, special attention should be made in choosing the localization radius and forget factor in data assimilation applications.

Key words: Lorenz96 model; LESTKF ; PDAF