王金成，李建平. 2010. 4DSVD 分析误差与样本选取方法和样本容量的关系初探［J］. 气候与环境研究，15（6）：729－742. Wang Jincheng，Li Jianping. 2010. Preliminary research on the relationship between the analysis error and both sampling strategies and sample content of the assimilation by using 4DSVD［J］. Climatic and Environmental Research (in Chinese)，15（6）：729－742.

4DSVD 分析误差与样本选取方法和样本容量的关系初探

王金成[1,2] 李建平[2]

1 国家气象中心，北京 100081
2 中国科学院大气物理研究所大气科学和地球流体力学数值模拟国家重点实验室，北京 100029

摘 要 分析误差与样本选取方法和样本容量的关系是 4DSVD 同化方法一个亟需研究的重要问题。获得支撑大气模式空间和观测空间吸引子的基向量是 4DSVD 研究的关键部分，样本的好坏和样本容量的范围是决定 4DSVD 基向量和分析结果质量的一个重要前提条件。首先利用 Lorenz 28 变量模式，用 4DSVD 方法做了一些简单三维同化试验，探讨了 Lorenz 28 变量模式的分析误差与样本容量和样本选取方法的关系。数值试验结果表明，对一个具体的模式，有限的样本容量就能够获得较高精度的分析结果；在模式系统和观测系统不变情况下，用一定样本容量得到的支撑模式空间和观测空间的基向量具有很好的稳定性，即一旦获得一组较好的基向量，在观测系统和模式系统不变的情况下，对同化任何时刻的观测适用；分析结果对选取方法没有太大的依赖性，但具体的样本容量要视不同模式和样本选取方法而定。用 WRF 模式做的 4DSVD 四维观测系统模拟试验结果表明，若样本选取方法得当，所需要的样本容量要远远小于模式自由度。4DSVD 要真正获得较高精度的分析结果，需要的条件是尽可能的在吸引子上取样并选取充足的样本容量；间隔取样可以一定程度上减少计算量。根据数值试验结果提出了 4DSVD 在实际同化时样本选取的一些初步的方法。

关键词 资料同化 4DSVD Lorenz 28 变量模式 WRF 样本选取方法 样本容量
文章编号 1006－9585（2010）06－0729－14 **中图分类号** P435 **文献标识码** A

Preliminary Research on the Relationship between the Analysis Error and Both Sampling Strategies and Sample Content of the Assimilation by Using 4DSVD

WANG Jincheng[1,2] and LI Jianping[2]

1 *National Meteorological Center, Beijing* 100081
2 *State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing* 100029

Abstract How to select samples used to obtain base vectors, and how many samples are enough for the data assimilation method based on singular value decomposition (4DSVD) are two important issues of the 4DSVD which need to be studied. Obtaining good base vectors which span the observation and model phase spaces of the attractor is a crucial part of the 4DSVD. And the quality of the base vectors is dependent on the selecting sample method and

收稿日期 2009－06－03 收到，2010－10－28 收到修定稿
资助项目 国家重点基础研究发展计划项目 2010CB950400，国家自然科学基金项目 40821092、41005055
作者简介 王金成，男，1981 年出生，博士，工程师，主要从事资料同化研究。E-mail：wangjc@cma.gov.cn
通讯作者 李建平，E-mail：ljp@lasg.iap.ac.cn

the sample content. Then the sampling strategies and sample content are important for 4DSVD. Some simple three-dimensional data assimilation numerical experiments about the relationship between the analysis error and both sampling strategies and samples content using 4DSVD are done, where Lorenz's 28-variable model is used. The numerical experiments results illustrate that limited samples could get good analysis field for a special problem. With certain conditions, base vectors which span the attractor space obtained from some special samples are stable, which implies that these base vectors could be used to assimilate data for similar problems at any time. So much computation time could be saved. Some results demonstrate that the analysis error is independent of the selecting sample method. The sample content which is suitable for 4DSVD depends on the special problem and the sampling strategies. The results of the observation system simulation experiment with WRF model illustrate that the necessary sample content is much smaller than the degree of the model freedom.

To get good analysis fields when do data assimilation using 4DSVD, the model is integrated from suitable initial conditions for a long time, a lot of samples which are enough and could represent the attractor are selected and the base vectors through SVD are obtained. Once the good base vectors are obtained, they can be used at any time for the similar problems. All of these results and conclusions are certainly helpful to the practical application of 4DSVD.

Key words data assimilation, 4DSVD, Lorenz's 28-variable model, WRF, sampling strategies, sample content

1 引言

4DSVD 方法是 Qiu and Chou（2006）基于大气吸引子理论（丑纪范，1986；李建平和丑纪范，1997；Li and Wang，2008）提出的一种新的同化方法，此方法将资料同化问题的解限定在吸引子相空间上，大大降低了计算量。研究表明，4DSVD 可以得到较高精度的分析场。这种方法最初提出的时候采用的是经验正变函数（EOF）方法从模式样本中获得支撑吸引子的基向量，但数值试验表明 SVD 方法作为获得支撑大气吸引子基向量的方法得到的分析场较 EOF 方法得到的分析场稳定（王金成等，2008）。4DSVD 基于吸引子理论，有着坚实的理论基础，是很有发展前景的一种资料同化方法。但 4DSVD 在具体操作和应用中还有一些亟待研究的问题，主要是分析误差与基向量个数的关系及分析误差与样本选取方法和样本容量的关系问题。误差与基向量的关系是 4DSVD 方法研究的首要问题，王金成等（2008）指出，4DSVD 的分析误差随基向量个数的增加而减少，当基向量个数达到一定数目时，分析误差达到最小，当基向量继续增多时，分析误差随基向量个数增加而增大；研究（Qiu et al.，2007a，2007b；Tian et al.，2008；Wang and Li，2009）指出，最优基向量个数还与观测误差和观测量个数有关。目前对 4DSVD 分析误差与样本选取方法和样本容量的关系研究还很少，本文对分析误差与样本选取方法和样本容量的关系进行了初步探讨。

2 模式简介

本文用的模式是 Lorenz（1965）在准地转两层大气模式基础上推导出来的 28 变量模式。Reinhold and Pierrehumbert（1982）应用和 Lorenz（1965）相似的 14 个正交函数（表 1），将准地转两层大气模式所有变量用这 14 个正交函数展开，然后得到了关于谱系数的常微分方程组：

$$\begin{cases}
\dfrac{\mathrm{d}\theta_i}{\mathrm{d}t} = \dfrac{1}{2}\sum_j\sum_k c_{ijk}(\theta_j\psi_k - \psi_j\theta_k) + \sigma_0\tau_i + h(\theta_i^* - \theta_i), \\
\dfrac{\mathrm{d}\psi_i}{\mathrm{d}t} = \dfrac{1}{2}\sum_j\sum_k c_{ijk}\left[(a_j^2 - a_k^2)(\psi_j\psi_k - \theta_j\theta_k) + \right. \\
\qquad \left. \hbar_j(\theta_k - \psi_k) + \hbar_j(\theta_j - \psi_j)\right]a_i^{-2} + \\
\qquad \beta\sum_j b_{ij}\psi_j a_i^{-2} - k(\psi_i - \theta_i), \\
\dfrac{\mathrm{d}\tau_i}{\mathrm{d}t} = \dfrac{1}{2}\sum_j\sum_k c_{ijk}\left[(a_j^2 - a_k^2)(\psi_j\theta_k + \theta_j\psi_k) - \right. \\
\qquad \left. \hbar_j(\theta_k - \psi_k) + \hbar_k(\theta_j - \psi_j)\right]a_i^{-2} + \\
\qquad \beta\sum_j b_{ij}\psi_j a_i^{-2} - \tau_i a_i^{-2} + k\psi_i - (k + 2k')\theta_i,
\end{cases}$$

其中，θ_i 为平均位温的谱展开系数，ψ_i 为平均流函数的谱展开系数，τ_i 为流函数切变的谱展开系数，i，j，k=1，2，…，14，其他参数的含义请参考

Reinhold and Pierrehumbert（1982）。

根据热成风关系，方程组（1）中变量 τ_i 转化成诊断变量，因此方程组剩下了 θ_i、ψ_i 的 28 个变量。由于这个谱模式首先由 Lorenz 推导得到，习惯上称这个谱模式为 Lorenz 28 变量模式。

表 1　Lorenz 28 变量谱模式的正交基函数
Table 1　Eigenfunctions of Lorenz's 28-variable Spectral Model

基函数	基函数
$F_1=\sqrt{2}\cos y$	$F_8=2\sin y\sin(2nx)$
$F_2=2\sin y\cos(nx)$	$F_9=2\sin(2y)\cos(2nx)$
$F_3=2\sin y\sin(nx)$	$F_{10}=2\sin(2y)\sin(2nx)$
$F_4=\sqrt{2}\cos(2y)$	$F_{11}=2\sin y\cos(3nx)$
$F_5=2\sin(2y)\cos(nx)$	$F_{12}=2\sin y\sin(3nx)$
$F_6=2\sin(2y)\sin(nx)$	$F_{13}=2\sin(2y)\cos(3nx)$
$F_7=2\sin y\cos(2nx)$	$F_{14}=2\sin(2y)\sin(3nx)$

Reinhold and Pierrehumbert（1982）指出 Lorenz 28 变量的关键参数是 θ_1^*，这个参数直接决定了模式的性质。当 θ_1^* 取一定的值时此模式是混沌系统，存在模式吸引子。有关 Lorenz 28 变量模式的具体参数和动力性质，请参考 Lorenz（1965）、Reinhold and Pierrehumbert（1982）。Lorenz 28 变量模式能很好的模拟大气的主要特征，因此被许多学者用来研究大气的动力性质和大气可预报性（Lorenz，1965；Reinhold and Pierrehumbert，1982；Krishnamurthy，1993）。

本文中 $\theta_1^*=0.15$，其他参数与 Reinhold and Pierrehumbert（1982）的参数相同，此时模式是混沌系统并且存在吸引子。为了达到一定的精度，差分方案选择精度较高的四阶龙格库塔格式，时间步长取为 0.01。

3　4DSVD 方法简介

4DSVD 同化方法首先由 Qiu and Chou（2006）提出，给出了理论公式，并建议采用 EOF 方法作为获得支撑大气吸引子相空间的方法。随后研究表明（王金成等，2008）采用 EOF 方法获得支撑大气吸引子基向量方法得到的分析结果不稳定，并且提出了用 SVD 方法作为获得支撑大气吸引子相空间基向量的方法，给出了具体的公式推导，并且得到了较好的分析结果。如无说明，下文 4DSVD 方法均指用 SVD 方法获得支撑大气吸引子基向量的 4DSVD 同化方法。

首先从模式积分结果中用一定的样本选取方法选取 N 个模式状态作为模式样本，其形成的矩阵记为 $\boldsymbol{S}$；通过观测算子 $\boldsymbol{H}$ 获得观测空间的模拟观测样本矩阵 $\boldsymbol{Z}$，模式样本和模拟观测样本分别表示如下，

$$\begin{aligned}\boldsymbol{S}&=(\psi_i)_{m,N},\\ \boldsymbol{Z}&=(\boldsymbol{H}(\psi_i))_{p,N},\end{aligned}\tag{2}$$

其中，m 和 p 分别是模式空间的维数和观测空间的维数。公式所用到的符号含义见表 2。

表 2　符号及含义
Table 2　Notations and meanings

数学符号	含义
$\boldsymbol{\psi}_\mathrm{t}$	准确的模式状态向量（m）
$\boldsymbol{\psi}_\mathrm{f}$	预报的模式样本向量（m）
$\boldsymbol{\psi}_\mathrm{a}$	分析向量（m）
$\boldsymbol{d}_\mathrm{o}$	同化时间窗内实际观测向量（p）
$\boldsymbol{d}_\mathrm{t}$	准确的观测向量（p）
$\boldsymbol{d}_\mathrm{fsim}$	观测空间的模拟观测向量（p）
$\boldsymbol{H}$	观测算子（$m\times p$）

对公式（2）给出的两个矩阵的协方差矩阵进行 SVD 分解：

$$\boldsymbol{S}\boldsymbol{Z}^\mathrm{T}=\boldsymbol{V}_\mathrm{L}\begin{pmatrix}\boldsymbol{E}&0\\0&0\end{pmatrix}\boldsymbol{V}_\mathrm{R},\tag{3}$$

其中，$\boldsymbol{E}$ 是对角矩阵，对角元素是奇异值；$\boldsymbol{V}_\mathrm{L}$ 和 $\boldsymbol{V}_\mathrm{R}$ 分别是模式空间和观测空间的奇异向量构成的矩阵（L 和 R 分别代表模式空间和观测空间），可以分解成如下形式：

$$\boldsymbol{V}_\mathrm{L}=(\boldsymbol{V}_{\mathrm{L}1},\boldsymbol{V}_{\mathrm{L}2},\cdots,\boldsymbol{V}_{\mathrm{L}r}),\tag{4}$$

$$\boldsymbol{V}_\mathrm{R}=(\boldsymbol{V}_{\mathrm{R}1},\boldsymbol{V}_{\mathrm{R}2},\cdots,\boldsymbol{V}_{\mathrm{R}r}),\tag{5}$$

其中 r 是理论上同化所采用的基向量的个数。设吸引子的维数是 s，根据 Whitney 定理，吸引子能够嵌入 $\boldsymbol{R}^{2s+1}$ 空间（Zhang and Chou，1992），即 $r=2s+1$，因此选取前（$2s+1$）个奇异向量作为支撑大气相空间吸引子的基向量。因此任何时刻的模式大气状态向量都可以用左奇异向量的线性组合表示，任何时刻的观测大气状态都可以用右奇异向量的线性组合来表示。

设 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 分别是模式空间和观测空间奇异向量的组合系数。对 N 个样本，可以求得每对奇异向量时间系数之间的关系是：

$$\boldsymbol{a}_{kt} = \rho_k \boldsymbol{b}_{kt}, \tag{6}$$

其中 ρ_k 是第 k 对奇异向量时间系数之间的线性系数。

观测资料可以用观测空间中奇异向量的线性组合来表示：

$$\boldsymbol{d}_o \approx \boldsymbol{V}_R \boldsymbol{x}, \tag{7}$$

用最小二乘法求得观测场的线性组合系数 $\boldsymbol{x}$，所以根据公式（6）中给出的左、右奇异向量组合系数的关系得到分析向量在模式空间各基向量的线性组合系数：

$$\boldsymbol{y}_k = \rho_k \boldsymbol{x}_k, \tag{8}$$

最后得到分析场的表达式：

$$\boldsymbol{\psi}_a = \boldsymbol{V}_L \boldsymbol{y} = \boldsymbol{V}_L \boldsymbol{\Omega} \boldsymbol{V}_R^T \boldsymbol{d}_o, \tag{9}$$

其中 $\boldsymbol{\Omega}$ 是由 ρ_k 为对角元素构成的对角矩阵。

4 数值试验

4.1 简单数值试验

4.1.1 试验设计

采用 Lorenz 28 变量模式，用随机的初值对模式进行积分，选取模式状态稳定后的一个时刻的值作为试验用初值。初始时刻 $t=0$ 变量 θ_i、ψ_i（$i=1, 2, \cdots, 14$）的初值如表 3 所示。设在 $t \in [1, 200]$ 中观测的时间间隔是单位时间，因只讨论 4DSVD 方法进行三维同化的情况，故单独对每个时刻的观测进行同化。用上述初值在时间 $t \in [1, 200]$ 的积分结果作为参考状态，在每个观测时刻的参考状态下叠加上一定方差的白噪声的误差，生成观测场，每个变量在相应的观测时刻都有观测。假设每个变量观测误差的均方差是该变量时间系列的均方差的 10%，具体的均方根误差如表 4 所示。

为了研究样本容量和样本选取方法对分析结果的影响，设计了如下 4 组试验（分别为 E1、E2、E3 和 E4），需要指出的是生成样本的初始场与生成参考状态的初始场是不相同的。

E1：采用表 3 给出的初值并叠加上方差为 0.001 的白噪声，产生两个不同的初始场，然后同时用这两个初始值积分模式，每个时间步长作为一个样本，称这样的取样方法为连续取样。积分时间步数 $N_{step} = 10, 20, \cdots, 100, 200, \cdots, 1000, 2000, \cdots, 5000$。

由于是由两个样本独立产生的，所以样本容量 $N = 2N_{step} = 20, 40, \cdots, 200, 400, \cdots, 2000, 4000, \cdots, 10000$。

表 3 模式初始值

Table 3 Initial values for the model

变量	初值	变量	初值
θ_1	1.0286700×10^{-1}	ψ_1	1.216330×10^{-1}
θ_2	-1.5415462×10^{-2}	ψ_2	-1.2547311×10^{-2}
θ_3	-3.7551325×10^{-2}	ψ_3	-5.9603665×10^{-2}
θ_4	-1.6524408×10^{-2}	ψ_4	-5.4365635×10^{-2}
θ_5	-1.2602359×10^{-2}	ψ_5	-2.2467805×10^{-2}
θ_6	8.8830460×10^{-3}	ψ_6	1.1973388×10^{-3}
θ_7	2.4508929×10^{-3}	ψ_7	1.5409050×10^{-2}
θ_8	-2.4389173×10^{-3}	ψ_8	-6.0133403×10^{-3}
θ_9	-9.4677750×10^{-3}	ψ_9	-2.5129095×10^{-2}
θ_{10}	1.4670164×10^{-2}	ψ_{10}	2.2857990×10^{-2}
θ_{11}	2.5653532×10^{-3}	ψ_{11}	3.4101950×10^{-3}
θ_{12}	-4.7655143×10^{-3}	ψ_{12}	-9.0692025×10^{-3}
θ_{13}	4.2177862×10^{-3}	ψ_{13}	6.8478780×10^{-3}
θ_{14}	6.5494678×10^{-3}	ψ_{14}	9.1685141×10^{-3}

表 4 观测误差的均方根误差

Table 4 Root-mean-square errors of the observation

观测变量	均方根误差	观测变量	均方误差根
θ_1	2.6831990×10^{-3}	ψ_1	1.0510765×10^{-2}
θ_2	8.3921859×10^{-3}	ψ_2	1.5914256×10^{-2}
θ_3	7.7989884×10^{-3}	ψ_3	1.4332153×10^{-2}
θ_4	5.1801992×10^{-3}	ψ_4	1.1581707×10^{-2}
θ_5	5.3149574×10^{-3}	ψ_5	1.0089772×10^{-2}
θ_6	5.2070934×10^{-3}	ψ_6	1.0138845×10^{-2}
θ_7	3.7500230×10^{-3}	ψ_7	7.6005203×10^{-3}
θ_8	3.7620305×10^{-3}	ψ_8	7.7062920×10^{-3}
θ_9	2.4655373×10^{-3}	ψ_9	4.4974689×10^{-3}
θ_{10}	2.5177200×10^{-3}	ψ_{10}	4.6866252×10^{-3}
θ_{11}	1.4505963×10^{-3}	ψ_{11}	2.4738745×10^{-3}
θ_{12}	1.4752549×10^{-3}	ψ_{12}	2.5069639×10^{-3}
θ_{13}	1.2091658×10^{-3}	ψ_{13}	1.8260564×10^{-3}
θ_{14}	1.2014614×10^{-3}	ψ_{14}	1.8484351×10^{-3}

E2：采用表 3 的初值对模式进行积分，从第 10000 步开始选取样本，同样是每个时间步长都作为一个样本，即连续取样。样本容量分别为 N ＝ 10，20，…，100，200，…，1000，2000，…，5000。

E3：采用表 3 的初值对模式进行积分，从第 50000 步开始选取样本，并且进行连续取样，样本容量分别是 N＝10，20，…，100，200，…，1000，2000，…，5000。

E4：采用表 3 的初值对模式进行积分，从第 50000 步开始选取样本，这次每间隔 10 个时间步长进行取样，这里称为间隔取样，样本容量和时间间隔分别是 N＝10，20，…，100，200，…，1000，2000，…，5000 和 ΔT_{step}＝10。

4.1.2　结果分析

采用 4DSVD 同化方案按照上述 4 个试验方案进行试验，为了衡量分析结果的好坏，定义相对误差，即分析场的均方根误差与观测均方根误差之比：

$$\varepsilon_{\text{a}} = \frac{\sqrt{\dfrac{\sum_{j=1}^{m}(\varepsilon_{\text{a}j})^2}{m}}}{\sqrt{\dfrac{\sum_{j=1}^{m}(\varepsilon_{\text{o}j})^2}{m}}}, \tag{10}$$

其中，ε_{aj}是每个变量的分析误差，即变量的分析值和参考值的差；ε_{oj}是每个变量的观测误差；m 是所有变量个数，对于 Lorenz 28 变量模式 m＝28。计算每个分析时刻分析场的相对误差，在本小节中如无特别说明，分析误差所指就是公式（10）定义的相对误差。

图 1 是试验 E1 所有分析时刻的平均分析误差与样本容量和基向量个数的关系图。从图 1a 可以看出，在样本数多于 5000，基向量个数在 17～21 个之间，存在着一个平均分析误差极小中心，说明样本容量为 5000 时，样本已经充足。图 1b 表明对于不同样本容量的试验，样本数小于 180 的试验中（除了样本个数为 20、40、60、80 的试验），平均分析误差随着基向量个数增大而减小，直到基向量个数为 28 的时候，平均相对误差都是 1，即分析误差和观测误差基本相当，分析没有改进模式初始场，也就是分析没有任何意义。这表明这种样本选取方法，样本容量小于 180 是不能够得到很好的基向量信息的。在样本容量为 20、40、60、80 的情况下，平均分析误差最初随着基向量的增加而减小，但当基向量增加到一定程度时，平均分析误差开始增大，所有的平均相对误差都大于 1。样本容量为 20 时，原因是样本容量小于模式空间的维数，加之样本间独立性不好，得不到足够的基向量；当样本容量为 40、60、80 时，可能原因是采用试验 E1 取样方法，样本间的不相互独立，样本空间的维数小于模式空间实际维数，由样本得到的基向量不足以支撑起模式空间。对于样本数大于 180 的所有试验，平均分析误差随着基向量个数增加而减小，直到分析误差达到最小（称此时的基向量个数为最优的基向量个数），并且所有最小的分析误差都小于 1，随后平均分析误差随着基向量个数的增加而增加，当基向量个数等于模式维数时，平均相对误差等于 1，这是由观测误差引起的，增加新的基向量引入了新的误差。这个结果与王金成等（2008）的研究结果一致。当样本达到一定数目后，最优基向量个数基本相同，这是因为在样本容量足够的情况下，4DSVD 得到的基向量能够很好的支撑模式空间和观测空间。采取试验 E1 的样本选取方法，样本容量大于 100 时，在一定的基向量个数条件下，平均分析误差小于观测误差，并且样本容量越大，平均分析误差越小，但当样本容量达到 5000 左右时，分析误差随着样本容量的增加而不再有明显改进。

对于试验 E1，样本容量少时最小分析误差较大，但当样本容量达到 5000 左右时，增加样本对分析场的改进作用就不大了，即此时样本已经足以得到支撑模式相空间的基向量了；在样本容量有限的情况下，分析误差达到最小的基向量个数与样本容量有关，也和观测误差大小有关，但当样本容量十分充足的情况下，所需要的最优基向量个数将不再随样本容量变化而变化。

从图 2a 中可以看出在样本容量达到 2000 以上时，基向量个数为 14～21 之间时，平均分析误差最小，相对误差可以达到 0.7 以下，说明在此时 4DSVD 的分析效果明显，相对观测而言，显著改善了模式的初始场。平均分析误差与样本容量和基向量个数的关系与试验 E1 基本相同，这里不再赘述。

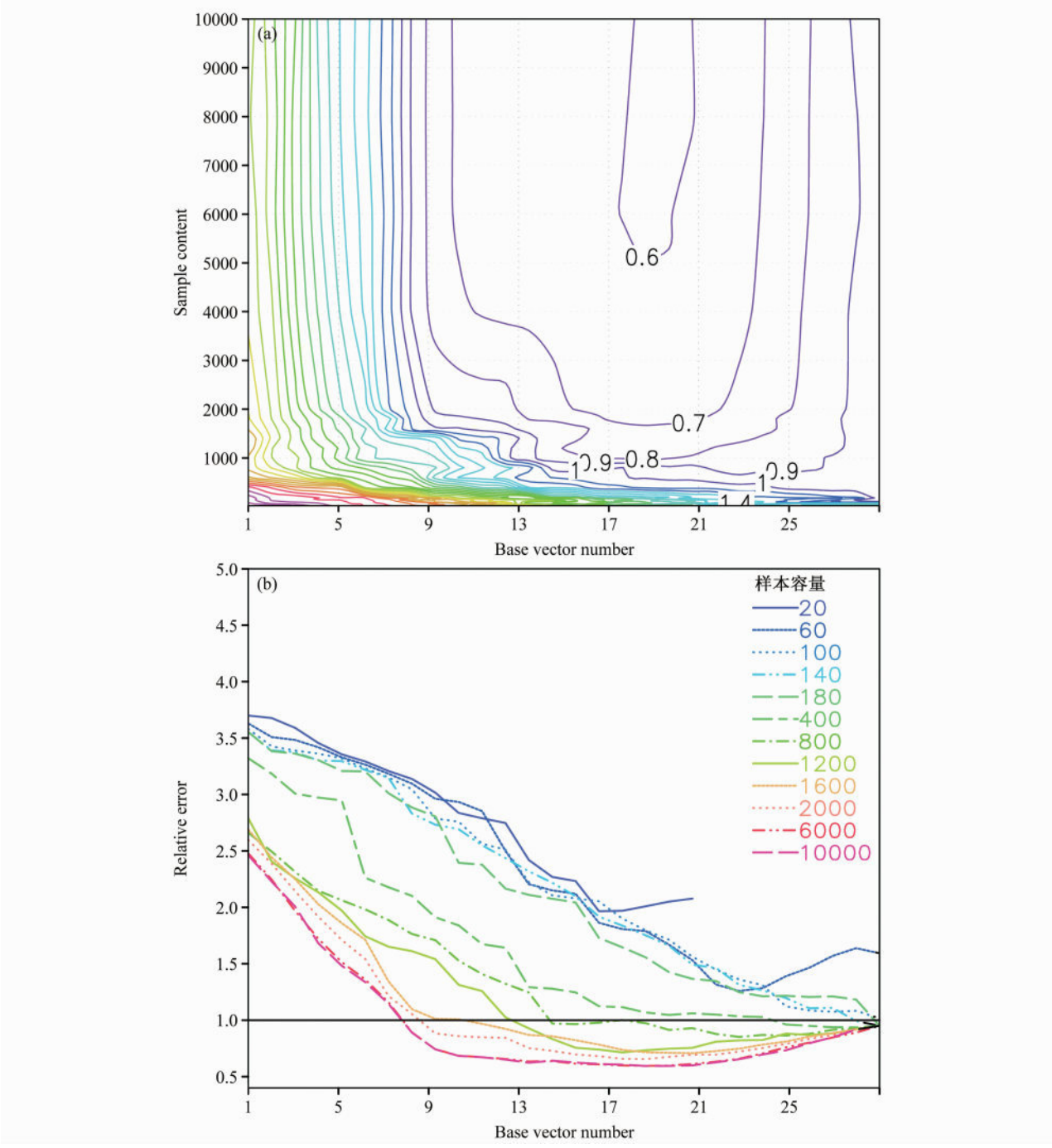

图 1 试验 E1（a）样本容量和（b）平均分析误差与基向量个数关系

Fig. 1 The relationships between both (a) sample content and (b) averaged analysis error and base vector number in experiment E1

从图 2b 可以发现在样本容量小于 100 时，平均分析误差开始随着基向量增多而减少，但极小的相对误差也大于 1，基向量继续增多时，平均分析误差开始增大，此时所有分析误差也都大于 1，这说明了此时样本容量不足，不能够得到支撑模式空间和观测空间的基向量。随着样本容量继续增多，最小平均分析误差在不断减小，最优基向量个数也在变化。当样本容量达到 2000 时，极小分析误差不再随样本增多而有明显减小，说明 2000 个样本已经提供了足够的信息，从而获得支撑模式空间的基向量。

比较图 2 和图 1，发现两种不同的取样方法，

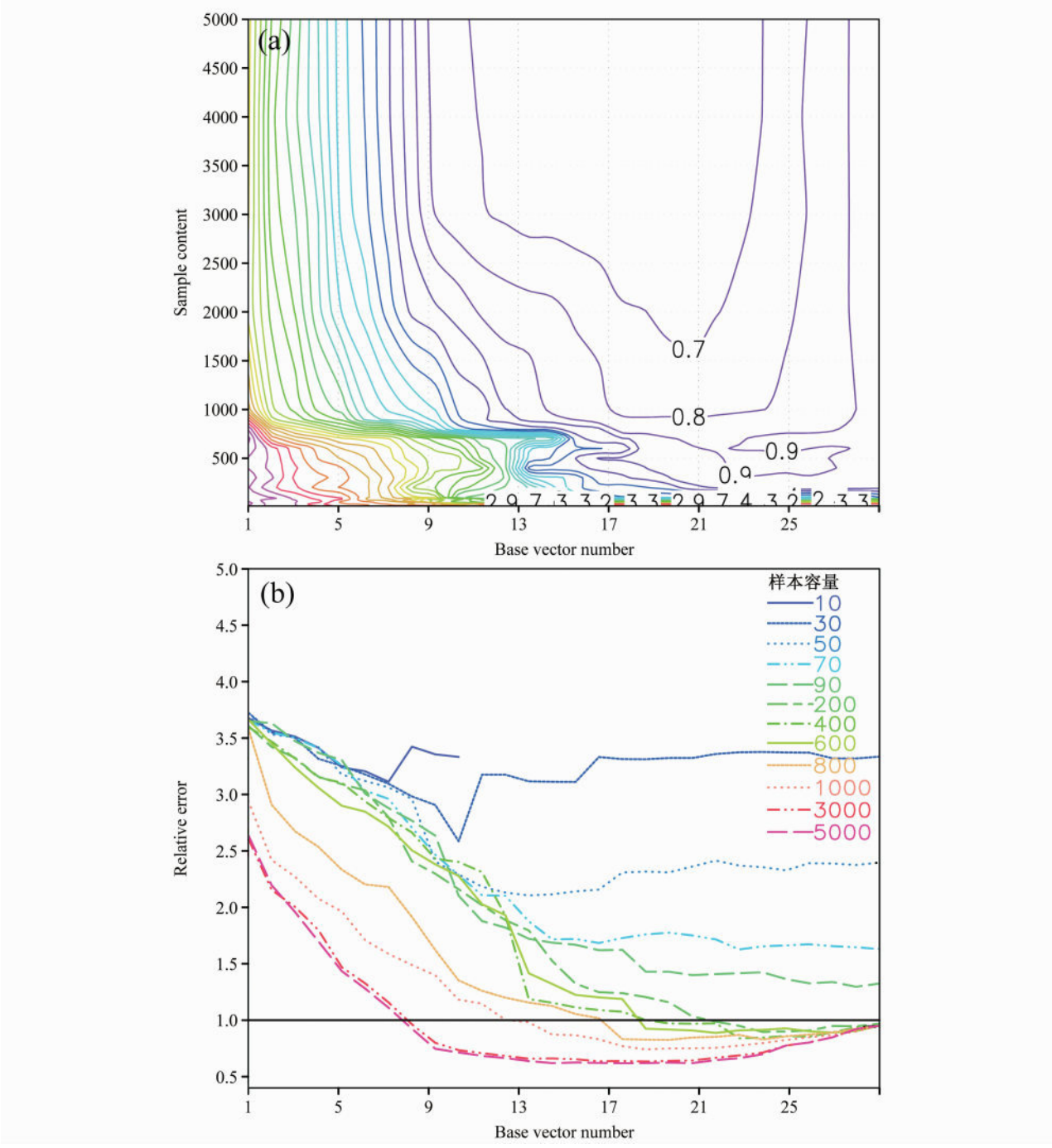

图 2　同图 1，但为试验 E2

Fig. 2　Same as Fig. 1，but for experiment E2

所得到的试验结果基本一致。说明采用不同的初值进行取样，在样本容量十分充足的条件下，分析误差在基向量个数达到一定数目，4DSVD 都能够明显改进模式初始场，且得到的最小分析误差基本相当。这说明 4DSVD 对取样的初值不敏感，只要保证充足的样本容量。

比较图 3 和图 2 发现，试验 E3 的分析误差与样本容量和基向量个数的关系特征与试验 E1、E2 结果基本相同。试验 E3 的分析误差在样本容量等于 100 左右、基向量个数小于 17 时，其随样本容量的增加迅速减小，递减率很大，而图 2a 中就没有如此明显。当样本容量大于 100 以上的区域，

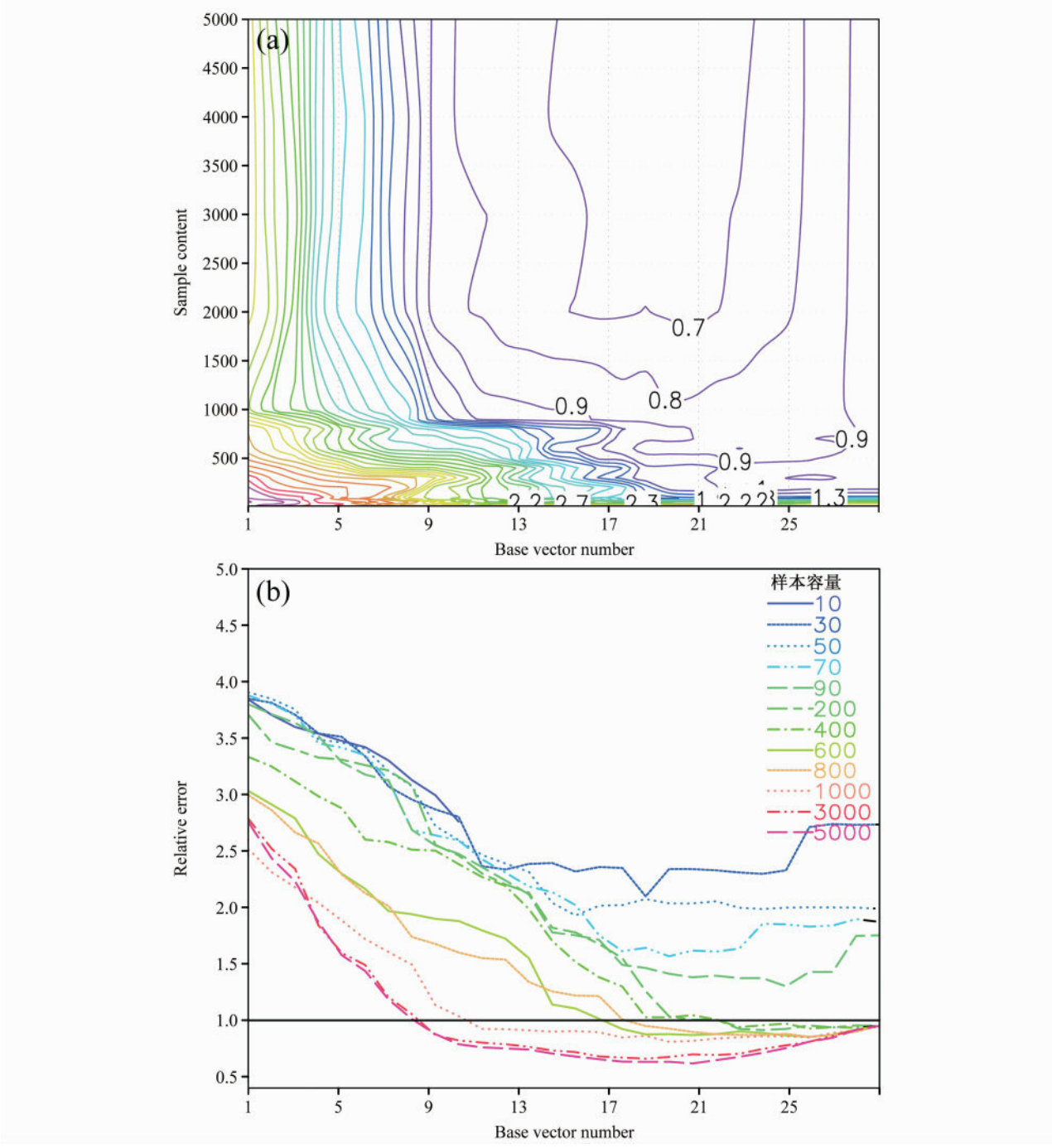

图 3 同图 1，但为试验 E3

Fig. 3 Same as Fig. 1, but for experiment E3

基向量个数为 20 左右时，随着样本的增多，分析误差迅速减小，当样本容量多于 2000 时，平均分析误差达到最小，随后平均分析误差随样本增多不再有明显减少。

比较图 1、图 2 和图 3，3 种方案中分析误差与基向量的关系基本完全相同，由于试验 E1 和试验 E2 与试验 E3 取得样本时的初值不同，试验 E2 和试验 E3 所选样本的区间段不同，因此当样本容量很充足时，分析误差对取样的区间段和得到样本的初始场不敏感。

对试验 E4（见图 4）而言，30 个样本就可以得到分析误差小于观测误差的分析场，而实验

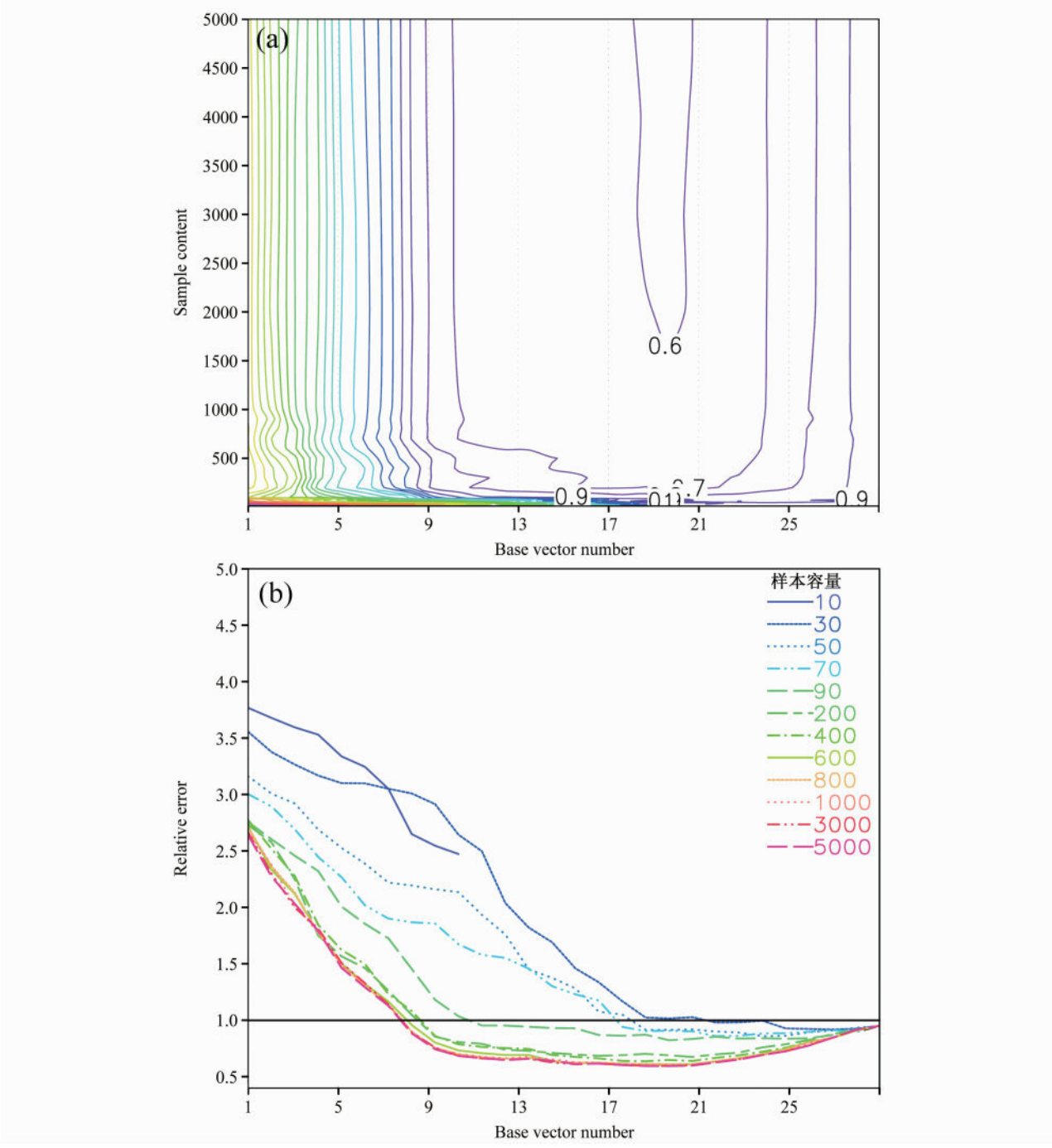

图 4　同图 1，但为试验 E4

Fig. 4　Same as Fig. 1，but for experiment E4

E1、E2 和 E3 需要 100 或者更多样本才能达到相同的效果。因此适度的间隔取样可以减少所使用的样本容量和计算量，因为间隔取样可以增加样本之间的独立性，使得样本之间的相关系数减少，代表性增大。

对上述 4 个试验其他时刻的分析误差（图略）进行分析，发现分析误差的特征以及分析误差与样本容量和基向量个数的关系基本相同，说明在 4 组试验不同试验条件下，同一种样本取样方法得到的支撑大气吸引子基向量具有稳定性，也就是说一种取样方法获得的基向量可以对同一模式的任何时刻的观测进行同化，大大地减少计算量；

样本容量很大时，分析结果不太依赖样本选择方法，不像集合卡曼滤波分析结果对初始样本集合具有很大的依赖性；不同的取样方法得到的样本虽然不同，获得的基向量也不尽相同，但它们最终的作用是相同的。4 组试验的分析误差达到最小所需要的样本个数分别为 5000、2000、2000、1500，最优基向量个数都为 20，说明不同的取样方法达到最小分析误差所需要的样本容量不同，间隔取样方案最优，达到最小分析误差所需要的样本最少，并且样本容量较少时，分析误差随样本增多而减少的速率最快，一定程度上减少计算量，因为间隔取样可以增加样本间的独立性，样本代表性增大；不论哪种取样方案，所需最优的基向量个数完全相同。

对试验 E4，在样本数小于 30 时，平均分析误差都大于 1，原因是样本容量不足造成的。而对试验 E1、E2 和 E3，在样本数分别大于模式自由度而小于 180、100 和 100 时，样本数已经远远大于模式自由度，但所得到的分析误差都要大于观测误差，当基向量比较大时分析误差还会增大。造成这种情况的原因是什么？可能是样本之间的独立性不够造成获得的基向量不好。为了验证此推论，图 5 给出了每组试验中所有试验的各个基向量解释方差。从图中可以看出，试验 E1、E2 和 E3 中样本容量分别小于 180、100 和 100 的时候，基向量解释方差很快减小到 0，非零基向量个数比较少，这说明所选样本之间的相互独立性太差。而试验 E4 在基向量大于 30 时，所获得的解

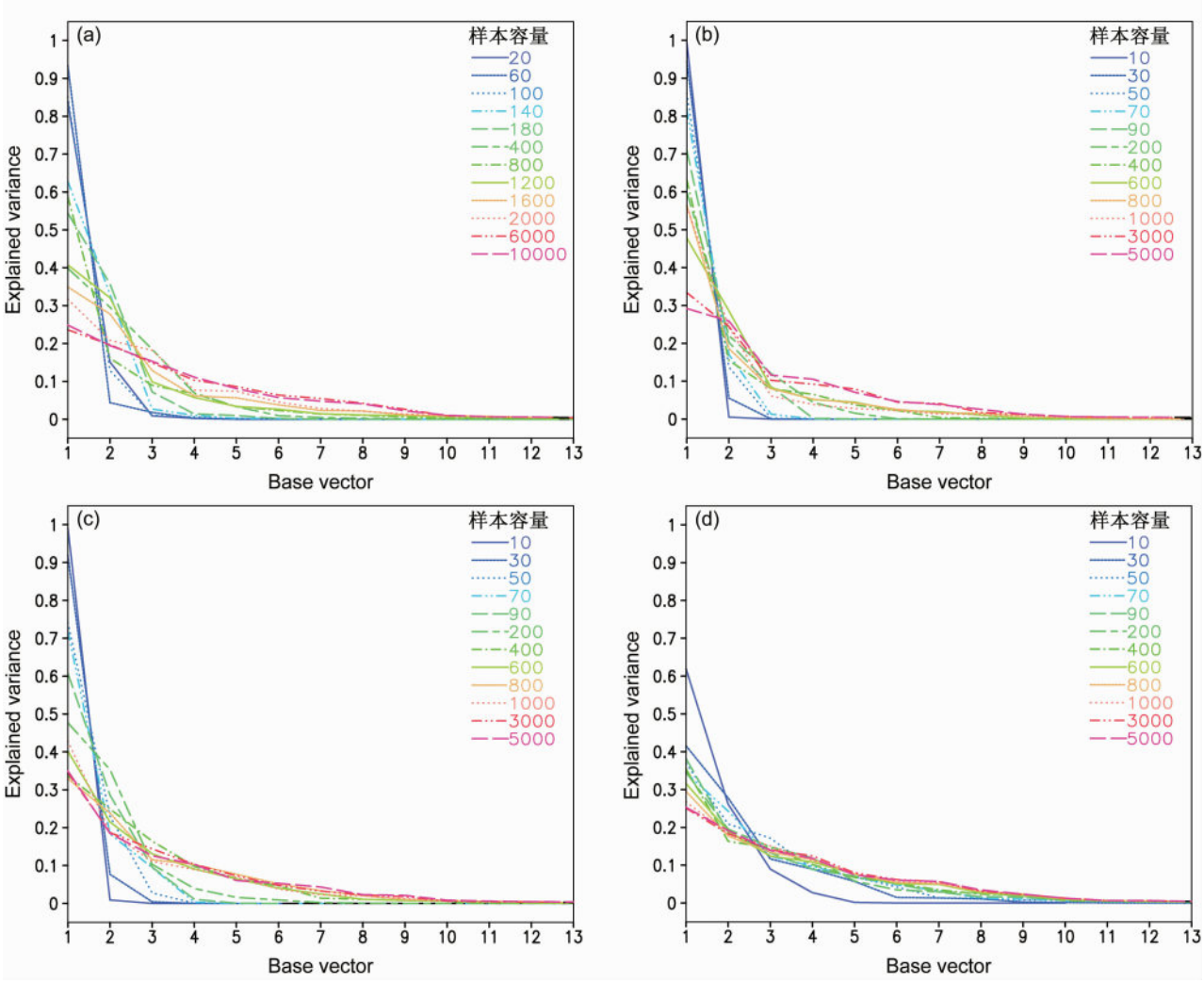

图 5 (a) 试验 E1、(b) 试验 E2、(c) 试验 E3 和 (d) 试验 E4 各基向量的解释方差

Fig. 5 The explained variance of base vectors for experiments (a) E1, (b) E2, (c) E3, and (d) E4

释方差大于零的基向量个数就明显多于试验 E1、试验 E2 和试验 E3。因此，造成试验 E1、E2 和 E3 样本数要远远大于模式自由度时才能获得较好的分析场的原因是所选取样本之间相互独立性不好，不能获得足够的非零解释方差的基向量。所以较好的取样方法可以减少样本使用量。

4.2　观测系统模拟试验

理论上，如果所选取样本间相互独立，根据 Whitney 定理，样本数只要大于 2 倍吸引子维数即可。由于实际大气模式吸引子维数远远小于模式自由度，那么获得较好基向量所需要的相互独立的样本数就远远小于模式自由度。上述 4 组试验的比较可以发现，间隔取样的 E4 比较好，建议实际应用中采用间隔取样方法，以保证样本相互独立。采用间隔取样方法，4DSVD 获得较好基向量所需要的样本容量远远小于模式自由度。为了验证这一结论，用 WRF 模式设计了观测系统模拟试验进行验证。观测系统模拟试验（Observation System Simulation Experiments，简称 OSSE）经常用来评估和检验计划中或者在建的观测仪器或观测系统作用（Miller，1990；Atlas，1997；Liu and Rabier，2003；Lahoz et al.，2005），也经常用来评估资料同化方法和系统（Kuo and Guo，1989；Bishop et al.，2001；Xue et al.，2006；Etherton，2007；Qiu et al.，2007b）。

4.2.1　试验设计

本小节中用到的模式是 ARW（Advanced Research WRF，简称 ARW）模式系统（Skamarock et al.，2005）。所设计的试验是初步的观测系统模拟试验。模拟区域是（25°N，112°E）。网格距设置为 30 km，水平网格为 40×60，垂直方向共 27 层，模式层顶为 50 hPa。所有试验中假设模式是完美的。初始场和边界条件均由 NCEP 最终分析资料(FNL)获得(https://dss.ucar.edu/datazone/dsszone/ds083.2[2008 - 11 - 04])。真值（参考值）是从 FNL 插值获得的初始场和边界条件下运行模式 24 h 后得到的。

本文模拟的是探空资料，变量包括温度和风速。方法是在上述参考状态上叠加高斯分布的白噪声。假设所有地面台站均有探空观测，位置分布如图 6。假设观测时间间隔是 12 h。观测的均方根误差的垂直扩线如图 7 所示。观测误差就是以图 7 中的数值为均方根的高斯白噪声。

样本选取方法如下：首先从 1998 年 1 月 1 日至 2007 年 11 月 30 日每隔一周选取一个时刻，用这些时刻的 FNL 资料作为初始条件和边界条件运行模式，选择 60 h 后的输出为样本，一共有 700 个样本，将这 700 个样本作为样本库。最后从这 700 个样本中随机抽取不等数量的样本进行试验，检验分析误差与样本容量的关系。

每选定一定数目的样本后，试验从 2007 年 12 月 1 日 00 时（北京时间，下同）开始，每 48 小时同化一次，共持续了 4 周，因此样本数一定的情况下，共有 14 个单独的同化试验。

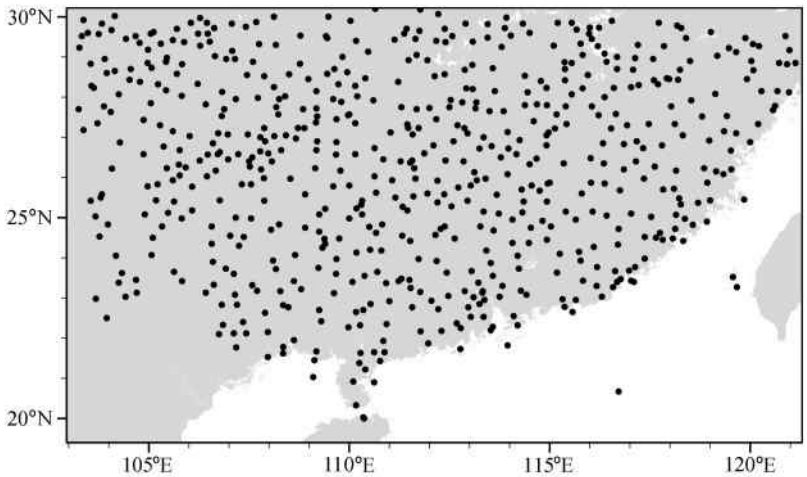

图 6　WRF 模式的模拟区域以及模拟观测网格（黑点）

Fig. 6　The WRF model domain and a snapshot of radiosonde observation network (black dots)

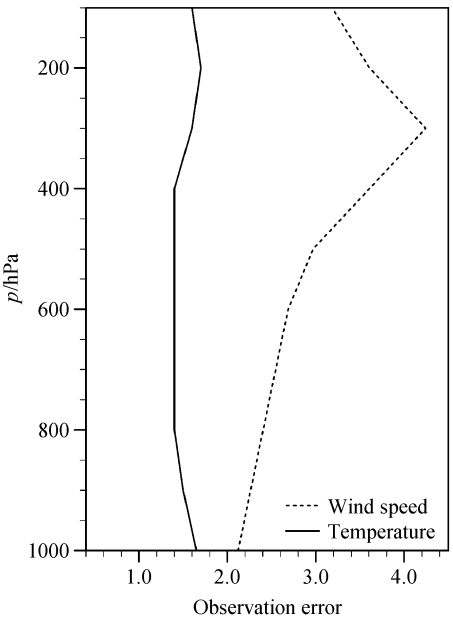

图 7 风速（单位：m·s^{-1}）和温度（单位：K）观测的均方根误差的垂直扩线

Fig. 7 Vertical profiles of the root-mean-square errors of observations for wind speed (m·s^{-1}) and temperature (K)

4.2.2 结果分析

以下结果所用的均方根误差都是在每个试验中的最小均方根误差。图 8 是不同样本容量试验中温度分析场均方根误差垂直廓线。从图中可以看出，各层平均均方根误差随着样本的增多而减少。样本数较小时，低层温度分析场均方根误差也要小于观测均方根误差；样本容量达到 100 时，各层分析场的均方根误差都小于观测场的均方根误差。图 9 是各层温度分析场均方根误差的平均值与样本容量的关系曲线。图 9 表明均方根误差随样本容量增多而减小，同时减小速率变慢。风场分析场的均方根误差与样本容量的关系与温度场相同（图略）。试验结果表明，在实际应用中间隔取样是很好的样本选取方法，同时获得较好的支撑模式空间的基向量需要的样本数远远小于模式自由度。

5 结论和讨论

本文讨论了 4DSVD 同化方法分析误差与样本

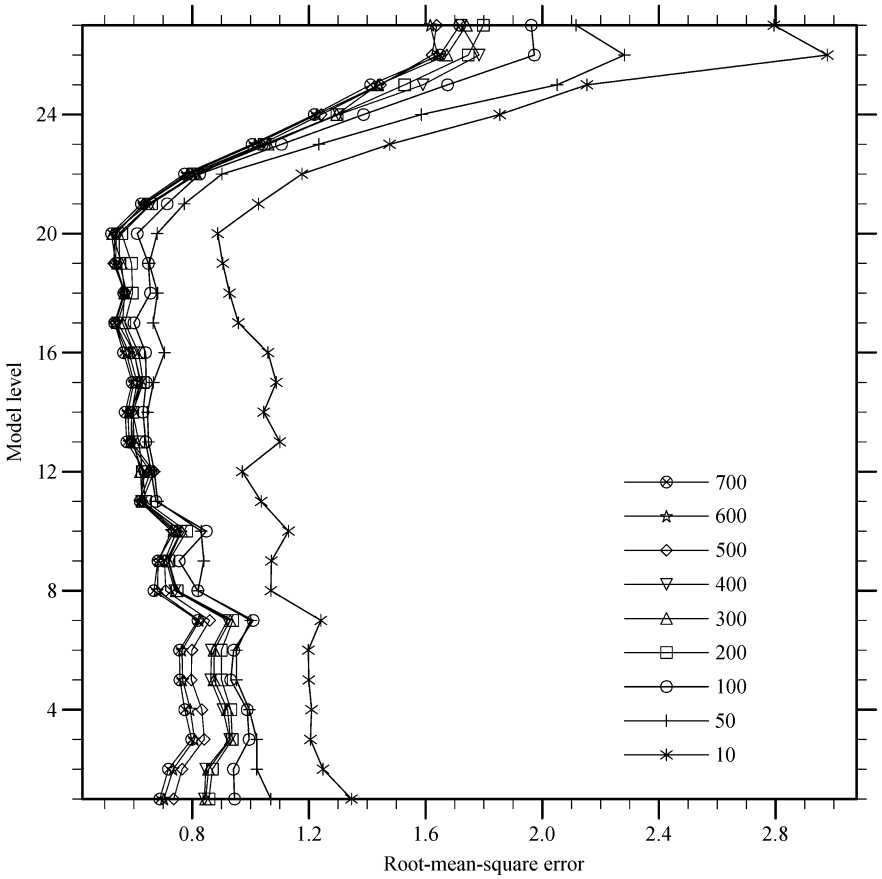

图 8 不同样本容量时温度分析场均方根误差垂直分布

Fig. 8 Vertical profile of the root-mean-square errors of analysis fields for temperature with different sample contents

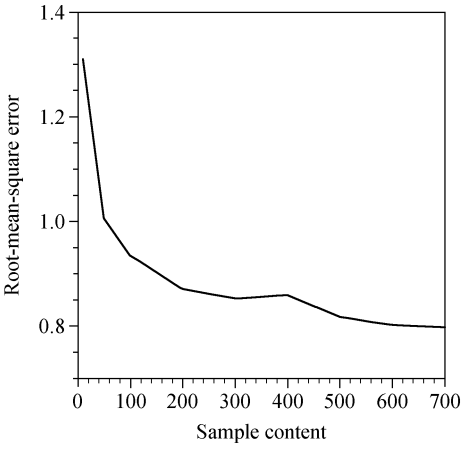

图 9　平均均方根误差与样本容量的关系

Fig. 9　The averaged root-mean-square error of analysis fields for temperature as the function of sample content

选取方法和样本容量的关系，同时用观测系统模拟试验对三维试验的部分结果进行了验证。

结果表明，4DSVD 方法在样本相互独立条件下，获得支撑模式空间基向量所需要的样本容量只要大于模式空间的维数即可；间隔取样是一个较好的样本选取方法，可以增大样本之间的相互独立性，可以一定程度的减少计算量；分析结果对选择样本的时间段和初值条件并不敏感；一旦获得了一组较好的基向量，这组基向量就可以确定下来进行同化任何时刻的观测场，即基向量具有稳定性。

4DSVD 的分析结果不但与基向量个数有关，还与样本容量有一定的关系，样本容量较少时，分析误差随样本容量增加而减少，但当样本容量达到一定时，分析误差不再随样本容量增加而减少，进入了稳定阶段，称此时的样本容量充足，即所选取的样本能够获得支撑相空间的基向量。这可以给以后应用 4DSVD 提供一定的指导，样本容量没有必要太多，只要“充足”即可。

分析误差与样本获取采用的初始条件无关，只要初始条件对模式来说合理即可，多个初始条件下获得样本与同一初始条件下获取样本只要样本容量相同，分析结果就基本相同。这说明在样本的选取过程中没有必要像集合卡曼滤波那样从一个初始样本集合去积分模式而获得样本，只要积分一定时间获得充足的样本容量即可。

适当间隔取样是一个比较好的选取样本的方法，间隔取样可以通过较少的样本得到与连续取样很多样本分析结果相同的基向量，在一定程度上减少计算量；在实际应用中采用间隔取样，获得较好的支撑模式空间的基向量所需的样本数要远远小于模式自由度。试验说明，充足的样本容量，合理的样本选取方法得到的基向量具有很好的实用性，就是说一旦获得了较好的基向量，就可以用它来同化模式结果的任意时刻，不必每次同化都要新的样本获得新的基向量。这不仅说明了得到的支撑大气吸引子的基向量好，还从另一个侧面表明了大气吸引子的存在，重要的是 4DSVD 方法可以大大降低计算量，可谓是“一劳永逸”。并且对于业务预报而言，历史预报已经积累了很多的预报大气状态，这是很重要的样本库。

根据试验结果和对试验结果分析，提出 4DSVD 取样的几个初步的方法：一是采用合理的初始场长时间积分模式（使得模式结果进入吸引子状态）；二是进行间隔取样，三是样本容量要很充足，但并不需要太多，需视具体情况而定。

本文探讨了 4DSVD 分析误差与样本选取方法和样本容量的关系，得到了一些有意义和启发性的结果。本文初步比较了几种样本选取方法的效果，结果表明，间隔取样的效果最好，那么如何定量选取基向量个数并给出一般的样本选取原则和方法将是实际应用 4DSVD 同化方法亟待解决的重要问题，这些都需要更加深入的研究。

参考文献（References）

Atlas R. 1997. Atmospheric observations and experiments to assess their usefulness in data assimilation [J]. J. Meteor. Soc. Japan，75：111－130.

Bishop C H，Etherton B J，Majumdar S J. 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I：Theoretical aspects [J]. Mon. Wea. Rev.，129：420－436.

丑纪范. 1986. 长期数值天气预报 [M]. 北京：气象出版社，329pp. Chou Jifan. 1986. Long-term Numerical Weather Prediction [M] (in Chinese). Bejing：China Meteorological Press，329pp.

Etherton B J. 2007. Preemptive forecasts using an ensemble Kalman filter [J]. Mon. Wea. Rev.，135：3484－3495.

Krishnamurthy V. 1993. A predictability study of Lorenz's 28-variable model as a dynamical system [J]. J. Atmos. Sci.，50 (14)：2215－2229.

Kuo Y H, Guo Y R. 1989. Dynamic initialization using observations from a hypothetical network of profilers [J]. Mon. Wea. Rev., 117: 1975-1998.

Lahoz W A, Brugge R, Jackson D R, et al. 2005. An observing system simulation experiment to evaluate the scientific merit of wind and ozone measurements from the future SWIFT instrument [J]. Quart. J. Roy. Meteor. Soc., 131: 503-523.

李建平, 丑纪范. 1997. 大气吸引子的存在性 [J]. 中国科学 (D辑), 27 (1): 87-96. Li Jianping, Chou Jifan. 1997. Existence of atmosphere attractor [J]. Science in China (Ser. D), 27 (1): 87-96.

Li J P, Wang S. 2008. Some mathematical and numerical issues in geophysical fluid dynamics and climate dynamics [J]. Commun. Comput. Phys., 3 (4): 759-793.

Liu Z Q, Rabier F. 2003. The potential of high-density observations for numerical weather prediction: A study with simulated observations [J]. Quart. J. Roy. Meteor. Soc., 129: 3013-3035.

Lorenz E N. 1965. A study of the predictability of a 28-variable atmospheric model [J]. Tellus, 17 (3): 321-333.

Miller R N. 1990. Tropical data assimilation experiments with simulated data: The impact of the tropical ocean and global atmosphere thermal array for the ocean [J]. J. Geophys. Res., 95: 11461-11482.

Qiu C, Chou J. 2006. Four-dimensional data assimilation method based on SVD: Theoretical aspect [J]. Theor. Appl. Climatol., 83: 51-57.

Qiu C J, Shao A M, Xu Q, et al. 2007a. Fitting model fields to observations by using singular value decomposition: An ensemble-based 4DVar approach [J]. J. Geophys. Res., 112, D11105, doi: 10.1029/2006JD007994.

Qiu C J, Zhang L, Shao A M. 2007b. An explicit four-dimensional variational data assimilation method [J]. Science in China (Ser. D), 50 (8): 1232-1240.

Reinhold B B, Pierrehumbert R T. 1982. Dynamics of weather regimes: Quasi-stationary waves and blocking [J]. Mon. Wea. Rev., 110 (9): 1105-1145.

Skamarock W C, Klemp J B, Dudhia J, et al. 2005. A description of the advanced research WRF version 2 [R]. NCAR technical note, NCAR/TN-468+STR.

Tian X J, Xie Z H, Dai A G. 2008. An ensemble-based explicit four-dimensional variational assimilation method [J]. J. Geophys. Res., 113, D21124, doi: 10.1029/2008JD010358.

王金成, 李建平, 丑纪范. 2008. 两种四维SVD同化方法的比较及误差分析 [J]. 大气科学, 32 (2): 277-288. Wang Jincheng, Li Jianping, Chou Jifan. 2008. Comparison and error analysis of two 4-dimensional singular value decomposition data assimilation schemes [J]. Chinese Journal of Atmospheric Sciences (in Chinese), 32 (2): 277-288.

Wang J C, Li J P. 2009. A 4DSVD scheme for Chaotic-Attractor-Theory oriented data assimilation [J]. J. Geophys. Res., doi: 10.1029/2008JD010916.

Xue M, Tong M, Droegemeier K K. 2006. An OSSE framework based on the ensemble square root Kalman filter for evaluating the impact of data from Radar networks on thunderstorm analysis and forecasting [J]. Journal of Atmospheric and Oceanic Technology, 23: 46-66.

Zhang B L, Chou J F. 1992. The application of empirical orthogonal functions to numerical simulation of climate [J]. Science in China (Ser. B), 35 (1): 92-101.