四维变分再分析输出的随机性

Xiaoqing Chen[1,2], Ross Bannister[3], Gavin Shaddick[1,2], 和 James V. Zidek [4]

[1]艾伦·图灵研究所；[2]埃克塞特大学数学与统计学系[3]雷丁大学气象系[4]不列颠哥伦比亚大学统计学系[1]*xchen@turing.ac.uk*

arXiv:2304.03648v1 [math.ST] 7 Apr 2023

摘要

本研究受欧洲中期天气预报中心的哥白尼大气监测服务（CAMS）再分析数据所启发。这些数据包含全球范围内最新的大气成分信息，是环境相关领域研究人员的重要资源。与通过监测设备获取的观测数据不同，此类再分析数据由计算机通过四维变分数据同化机制生成，因此其随机属性仍很大程度上尚不明确。对这一属性缺乏了解，进而限制了这些数据的适用范围，也妨碍了它们在统计学上的广泛、灵活应用，尤其是时空建模方面的应用；目前其用途主要局限于不确定性量化和数据融合。因此，本文研究了这些再分析输出数据的随机属性。我们运用测度论，证明了与这些再分析数据相关的空间和时间随机性确实存在，并揭示出这些数据本质上是现实世界中不可观测的空间和／或时间随机过程经数字化后得到的实现。这意味着在实际应用中，我们可以像处理观测数据一样处理再分析输出数据，从而使更灵活的时空随机方法能够应用于这些数据。我们还对再分析数据中的不同类型误差进行了客观分析，并阐明了这些误差之间的相依或独立关系，为误差项建模提供了明确、具体的指导。本研究的结果也为时空建模者和环境人工智能研究人员奠定了坚实基础，使他们能够直接利用这些再分析输出数据开展基于随机模型的研究。

关键词— 欧洲中期天气预报中心哥白尼大气监测服务再分析数据、四维变分再分析输出、时空过程、随机过程

1 引言

本研究受欧洲中期天气预报中心的CAMS再分析数据集启发。

欧洲中期天气预报中心（ECMWF）代表欧盟实施哥白尼大气监测服务（CAMS）（ECMWF，未注明日期），并制作 CAMS 再分析数据产品。该产品是最新的大气成分再分析数据集，涵盖全球范围内的气溶胶（如 PM2.5、PM10）、化学物质（如硫酸盐、黑碳）以及温室气体（如 CO2、NO2 等），包含四维（三个空间维度和一个时间维度）大气成分场，目前覆盖 2003年至2016年（此后每年向前延展一年）（Inness 等，2019）。

因此，欧洲中期天气预报中心哥白尼大气监测服务再分析数据产品已成为许多不同环境相关研究领域（如时空建模、公共卫生、气候变化和环境智能）研究人员的重要且信息丰富的数据来源。

等，以研究环境相关研究主题，并支持政策制定者应对环境相关问题及其相应影响。

与通常通过监测设备获得、因而通常被视为来自具有内在随机性的现实世界隐含空间和/或时间过程的一次实现的观测测量数据集不同，欧洲中期天气预报中心哥白尼大气监测服务再分析数据是通过一种 增量四维变分数据同化方法（Courtier等，1994年）生成的。在该方法中，总柱 CO、气溶胶光学厚度等卫星反演观测值通过从09:00到 21:00的 12 小时同化窗口，以及两种谱截断 T95（210 km）和 T159（110km）进行同化（Inness等，2019年），因此这些数据并非直接来自监测设备，而是由计算机获得的。

同样由计算机产生的传统数值天气预报模型通常使用偏微分方程来模拟大气成分的物理规律（Lorenz，1963年），因此其模型输出通常被认为是确定性的。然而，在当前文献中，通过计算机生成但经由四维变分数据同化（DA）机制产生的再分析输出究竟是确定性的还是随机性的，仍然很不明确。

因此，这类计算机生成的再分析输出仅适用于经验性用途，例如气候统计计算（如气候平均值、百分位数等）、趋势研究、地理可视化制图（Inness 等，2019），或有限的统计应用，其中一类是量化不确定性，另一类是数据融合。

量化不确定性的方法是，假设计算机执行的数学模型具有确定性，并将其视为一个“黑箱”，不利用模型中的任何数学信息（Kennedy 等，2001，第 1.1 节），同时坚持认为模型输出中存在多种不确定性。Kennedy 等（2001，第 2 节）提出了一种可能的方法，可在假定计算机模型为“黑箱”的前提下，对不同的不确定性来源进行主观分类。

显式量化方法是基于集合中的计算机模型输出构建统计模型。在确定性计算机模型输出的情况下引入统计模型的理由，是集合成员的可用性有限，这可能是由于每次运行的计算时间成本高昂、可用的不同模型有限、资金不足等原因造成的（Sain 等，2011，第 1 节）。

融合方法的核心思想是将计算机模型输出与观测数据联系起来，参见Kalnay（2003年，第5章）及其中的参考文献。近期发展起来的一种融合方法，尤其适用于时空研究领域，是降尺度模型（Berrocal等人，2010）。降尺度模型以线性回归的方式，将位置 s 处的监测观测数据 $Z(s)$与特定网格单元 B 的计算机模型输出 $x(B)$联系起来，同时加入加性高斯分布随机测量误差；随后，通过一种称为协区域化线性模型（LMC）的经典多元建模方案，将其随空间变化的回归系数进一步联合建模为二元高斯分布空间过程，参见Wackernagel（2013年，第172页）或Banerjee等人（2014年，第278页）。有关该方法的更多细节，请参见Berrocal等人（2010）或Zidek等人（2012）。有关将此类方法直接应用于CAMS再分析数据的研究，请参见Shaddick等人（2020）。

然而，这种融合建模方法与观测值在一定程度上涵盖了计算机输出数据的整体随机特性。例如，在 Shaddick等人（2020）中，读者无法确定降尺度模型中的高斯随机过程和随机误差究竟是完全由随机监测观测所致，还是由监测观测与计算机输出的CAMS再分析数据共同引起。

计算机生成的再分析数据具有模糊且未知的随机性，因此妨碍了此类数据集在更广泛、更灵活的统计建模应用中的实用性。例如，目前尚不清楚标准的随机时空建模框架（包含加性空间和时间随机项）是否适用于这类数据，例如，数据 $=$ 协变量 $+$ 空间随机效应 $+$ 时间随机效应 $+$ （时空交互作用） $+$

随机测量误差 （Cressie 等，2015年，第304–305页）是否能够仅应用于这些四维变分再分析输出（即欧洲中期天气预报中心的哥白尼大气监测服务再分析数据）；以及当特别关注该数据集中不同大气成分之间的联合关系时，是否能够仅使用这类CAMS再分析数据，而无需与其他监测观测数据结合，来应用经典的多元空间建模框架（例如随机协同克里金法）。

诚然，有人可能会从贝叶斯视角出发，认为一切都可以是随机的。我们的观点是，贝叶斯随机性只能被动地与已构建模型的各个部分（包括参数）相关联，却无法主动论证或指示模型结构是否应包含空间和/或时间随机成分；这几乎完全取决于数据集本身的性质和特征。只有当数据集确实呈现出空间和/或时间随机特性时，我们才能将这些成分纳入模型。或者，从一般误差的角度来看，只有当误差由空间和/或时间来源引起时，我们才能将简单回归模型的一般误差项分解为空间随机效应、时间随机效应和其他测量误差。

为此，本文研究了四维变分再分析输出的随机属性。据我们所知，我们首次揭开了四维变分资料同化机制的“黑箱”，并运用测度论证明了与再分析输出相关的时空随机性的存在。与 Kennedy 等人（2001年，第 2 节）主观假设计算机模型输出中可能存在误差不同，我们依据四维变分机制，客观分析了不同类型的误差来源及其相互依赖或独立关系，从而全面理解再分析输出的随机性，并因此能够明确回答上述实际建模过程中关于哥白尼大气监测服务再分析数据的不确定性问题。

在第2 节中，我们阐述四维变分资料同化机制，以揭开“黑箱”。在第3 节中，我们向读者介绍与随机过程相关的概念，并提出关于四维变分再分析输出中存在随机性的猜想。测度论证明位于第 4 节，在该节中，我们从两个不同的角度展开证明。在第5节中，我们重点剖析与这些再分析输出相关的不同类型的随机误差及其相互依赖性/独立性。在第 6 节中，我们通过讨论结束本文，其中包括本研究的实际意义。

2 四维变分数据同化的机制

在本节中，我们将“黑箱”四维变分资料同化具体化。

（Bannister，2001年）四维变分资料同化是一种为每个资料同化循环寻找最佳可能初始状态$\mathbf{x}_{\cdot}$的方法。优化后的初始状态$\mathbf{x}_{\cdot}^{A}$（此处下标中的点表示任意迭代循环）实际上事实上 就是我们所需的再分析输出，它通过最小化代价函数J 获得；该代价函数衡量两个项之间的偏差：一项是每个同化循环中的任意状态$\mathbf{x}_{\cdot}$ 与所谓的背景状态$\mathbf{x}_{\cdot}^{B}$之间的差异，另一项是在给定资料收集时间窗内的预报观测$\boldsymbol{H}_{\tilde{t}}^{o}\cdot\mathbf{x}_{\tilde{t}}$与在这一特定时间窗内收集的真实观测$\mathbf{y}_{\tilde{t}}$ 之间的差异。

代价函数J 被定义为：

$$J[\mathbf{x}_{k\Delta t}]=\frac{1}{2}(\mathbf{x}_{k\Delta t}^{B}-\mathbf{x}_{k\Delta t})^{T}\mathbf{B}_{k\Delta t}^{-1}(\mathbf{x}_{k\Delta t}^{B}-\mathbf{x}_{k\Delta t})+\frac{1}{2}\sum_{\tilde{t}=k\Delta t}^{(k+1)\Delta t}(\mathbf{y}_{\tilde{t}}-\mathbf{H}_{\tilde{t}}^{o}\cdot\mathbf{x}_{\tilde{t}})^{T}\mathbf{R}_{\tilde{t}}^{-1}(\mathbf{y}_{\tilde{t}}-\mathbf{H}_{\tilde{t}}^{o}\cdot\mathbf{x}_{\tilde{t}}),\ (1)$$

其中

- k：每次同化循环运行的计数器； $k=0$、1、2、.....；

- $\mathbf{x}_{k\Delta t}$：同化的（$k\Delta t$）th 运行对应的初始模式状态，它是一个向量，汇集了给定时间步 t（例如 $k=0$，然后是 $k\Delta t \triangleq t=0$）所有网格位置上的大气成分值，例如 PM2.5、黑碳（BC）、硫酸盐（SU）等。该初始状态的优化值记为$\mathbf{x}_{k\Delta t}^{A}$，即所需的再分析输出结果，正是由欧洲中期天气预报中心提供的 CAMS再分析产品；

- $\mathbf{x}_{k\Delta t}^{B}$：称为同化的（$k\Delta t$）th 次运行中的背景状态，它通过将数值天气预报（数值天气预报）模型（一系列时间算子的乘积）应用于上一次同化运行的再分析输出而获得，例如，$\mathbf{x}_{\Delta t}^{B}=\mathbf{M}_{\Delta t}\cdot\mathbf{M}_{\Delta t-\delta t}\cdots\mathbf{M}_{\delta t}\cdot\mathbf{M}_{0+\delta t}\mathbf{x}_{0}^{A}$，此处 δt 是窗口 $0\sim\Delta t$内的增量时间步长；

- $\mathbf{B}_{k\Delta t}$：包含给定同化时间下不同大气成分在不同网格位-置上的背景状态误差的方差和协方差的背景状态误差协方差矩阵 $t=k\Delta t$；误差可以是系统性的（例如偏差），也可以是随机的。偏差通常会在同化程序开始前进行校正，而随机误差则假定服从高斯分布；在制作CAMS再分析产品时，不同场之间的误差协方差矩阵是分块对角的，这意味着每个场都以单变量方式进行同化（Inness 等，2019，第二节3）；

- $\mathbf{y}_{\tilde{t}}$：在时间窗口 $\tilde{t}=k\Delta t\sim(k+1)\Delta t$ 内收集的观测；

- $\mathbf{x}_{\tilde{t}}$：时间窗口 $\tilde{t}=t\sim t+\Delta t$ 期间的模式状态$\tilde{t}=t$，通过数值天气预报模型从时间$\tilde{t}=t$ 演变获得，例如，如果 $\tilde{t}=0\sim\Delta t$，则 $\mathbf{x}_{\Delta t}=\mathbf{M}_{\Delta t}\cdot\mathbf{M}_{\Delta t-\delta t}\cdots\mathbf{M}_{2\delta t}\cdot\mathbf{M}_{0+\delta t}\cdot\mathbf{x}_{0}$；数值天气预报模型 $\mathbf{MM}\cdots\mathbf{M}$ 的详细信息见附录 A。

- $\mathbf{H}_{\tilde{t}}^{o}\cdot\mathbf{x}_{\tilde{t}}$：预报观测：通过将给定模式状态$\mathbf{x}_{\tilde{t}}$ 左乘观测位置平滑算子矩阵$\mathbf{H}_{\tilde{t}}^{o}$ 获得，该矩阵包含不同位置处不同观测的插值系数；有关$\mathbf{H}_{\tilde{t}}^{o}$ 的结构详情，见附录 B；

- $\mathbf{R}_{\tilde{t}}$：时间窗内的观测误差协方差矩阵 t；

- 注：下标用于表示离散时间步。

因此，基于上述内容，下面我们详细阐述欧洲中期天气预报中心哥白尼大气监测服务再分析数据产品的生成过程；该过程也为深入理解这些再分析输出及其相应误差的性质提供了绝佳机会。

在第一次运行（或时间步）中，$k=0$，因此 $k\Delta t\triangleq t=0$，以及 $\tilde{t}=0\sim\Delta t$，收集观测 $\mathbf{y}_{\tilde{t}}$，

- 背景状态向量$\mathbf{x}_{0}^{B}$在初始时刻 $t=0$ 被假定为一个猜测；
- 状态 $\mathbf{x}_{0}$ 在初始时刻（$t=0$）被设定为与上述 $\mathbf{x}_{0}^{B}$ 相同；
- 将模式状态$\mathbf{x}_{\tilde{t}}$ 从 $\tilde{t}=t=0$ 推进到窗口 $\tilde{t}=0\sim\Delta t$ 内所需的任意离散时间步，经过 $\mathbf{x}_{\tilde{t}}=\mathbf{M}_{\tilde{t}}\cdot\mathbf{M}_{\tilde{t}-\delta t}\cdots\mathbf{M}_{2\delta t}\cdot\mathbf{M}_{0+\delta t}\cdot\mathbf{x}_{0}$；
- 乘以观测位置平滑算子$\mathbf{H}_{\tilde{t}}^{o}$ 以获得模拟观测$\mathbf{H}_{\tilde{t}}^{o}\cdot\mathbf{x}_{\tilde{t}}$；

然后，通过相对于 $\mathbf{x}_{0}$ 最小化该次运行的代价函数 J，我们得到该次运行初始状态对应的第一个优化再分析结果，并将其记为$\mathbf{x}_{0}^{A}=\arg\min\ J$，以替代初始猜测$\mathbf{x}_{0}$。

在第二次运行中，$k=1$，因此 $t=\Delta t$，并且 $\tilde{t}=\Delta t\sim2\Delta t$，收集观测 $\mathbf{y}_{\tilde{t}}$，

- 本次运行的初始时刻为 $t = \Delta t$，因此背景状态向量为$\mathbf{x}_{\Delta t}^{B}$，这是将数值天气预报模型应用于上一次运行的再分析输出$\mathbf{x}_{0}^{A}$所得到的，即 $\mathbf{x}_{\Delta t}^{B} = \mathbf{M}_{\Delta t} \cdot \mathbf{M}_{\Delta t - \delta t} \cdots \mathbf{M}_{0+\delta t} \cdot \mathbf{x}_{0}^{A}$;

- 本次运行的模式状态为 $\mathbf{x}_{\tilde{t}(=\Delta t \sim 2\Delta t)} = \mathbf{M}_{\tilde{t}} \cdot \mathbf{M}_{\tilde{t}-\delta t} \cdots \mathbf{M}_{\Delta t+2\delta t} \cdot \mathbf{M}_{\Delta t+\delta t} \cdot \mathbf{x}_{\Delta t}$;
- 乘以观测位置平滑算子$\mathbf{H}_{\tilde{t}}^{o}$以获得本次运行的模拟观测$\mathbf{H} \cdot \mathbf{x}_{\tilde{t}}$；

并针对模式状态$\mathbf{x}_{\Delta t}$ 最小化代价函数 J，随后我们得到此次运行初始状态对应的第二个优化再分析结果，并将其记为$\mathbf{x}_{\Delta t}^{A} = \arg\min J$。

按照相同的逻辑，我们可以得到一组再分析输出$\{\mathbf{x}_{0}^{A}, \mathbf{x}_{\Delta t}^{A}, \mathbf{x}_{2\Delta t}^{A}, \ldots\}$。

代价函数 J 的示例推导，以及前两次运行的相应一阶导数表达式（针对arg minJ），详见附录 C，尤其是公式（7）。一般而言，时间 t 的再分析输出 $\mathbf{x}_t \triangleq \mathbf{x}_t^A$形式为

$$\mathbf{x}_t \triangleq \mathbf{x}_t^A \propto L\mathbf{x}_t^B + \sum_{\tilde{t}=t}^{t+\Delta t} K_{\tilde{t}} \mathbf{y}_{\tilde{t}}, \tag{2}$$

其中， L和 $K_{\tilde{t}}$是由$\mathbf{HM} \cdots \mathbf{M}$组成的两个系数矩阵。特别是，$K_{\tilde{t}}$ 中涉及$\mathbf{M}$ 的项数取决于 $\tilde{t}$。推导细节见附录 C 中的公式（8）和（9）。

3 再分析输出中存在随机性的猜想

本节提出我们关于上述再分析输出$\{\mathbf{x}_{0}^{A}, \mathbf{x}_{\Delta t}^{A}, \mathbf{x}_{2\Delta t}^{A}, \ldots\}$中存在随机性或随机现象的猜想，等价地说，也就是欧洲中期天气预报中心哥白尼大气监测服务的再分析数据中存在随机性或随机现象。

首先，我们简要回顾一些与随机过程相关的基本概念。我们遵循 Grimmett 等人（2001年，第 360 页）和 Shalizi （ 2007 年，第 3–5 页）所采用的惯例。

3.1 随机过程的概念

随机过程$stochastic\ process$ $\mathbf{X}$ 是随机变量 X_t或一般意义上的随机对象 [1] $\mathbf{X}_t$的一个集合或族$\{\mathbf{X}_t : t \in T\}$，其中每个变量或对象都将样本空间 Ω 映射到状态空间 U，即 $X_t : \Omega \to U$ 或$\mathbf{X}_t : \Omega \to U$。

指标集 T可以是离散的，例如 {0、1、2、.....}，也可以是连续的，例如 $[0, \infty)$；状态空间 U 可以是整数 $\mathbb{Z}$ 或实数 $\mathbb{R}$。指标集 T 和状态空间 U 的选择共同决定随机过程的解析性质。

对于任意固定的样本元素 $\omega \in \Omega$，都有一个对应的集合 $\{\mathbf{X}_t(\omega) : t \in T\} \subset U$，称为随机过程 $\mathbf{X}$ 在 ω处的一次实现或一条样本路径。

当索引集 T 的长度为 1 时，也就是说索引集 T 只有一个元素，那么随机过程就退化为平凡随机过程，或者说等价于一个随机变量。

[1]当输出空间的 $sigma$-代数属于 $\mathbb{R}$[1]时，它是一个随机变量；当输出空间的 $sigma$-代数属于 $\mathbb{R}^n$时，它是一个随机向量；当输出空间的 $sigma$-代数是一个序列时，它是一个随机过程。

3.2 随机性存在性猜想

现在回到四维变分资料同化的再分析输出集合 $\{\mathbf{x}_0^A, \mathbf{x}_{\Delta t}^A, \mathbf{x}_{2\Delta t}^A, \ldots\}$，根据方程（1）和（2），我们知道，每次运行 t 时的再分析输出$\mathbf{x}_t^A$本质上是时刻 t 的背景状态$\mathbf{x}_t^B$以及从 t 到 $t+\Delta t$ 收集的观测$\mathbf{y}_{\tilde{t}}$的函数。也就是说$\mathbf{x}_t^A = \arg\min\ J = \mathfrak{G}(\mathbf{x}_t^B, \mathbf{y}_{\tilde{t}})$。

尽管背景状态$\mathbf{x}_t^B$是从上一次运行的再分析输出中获得的，其随机性仍有待证明（初始运行除外，在初始运行中$\mathbf{x}_t^B$被设为一个猜测值），因此目前可将其视为常量（更多讨论请见第 6 节）；但从 t 到 $t+\Delta t$ 收集的观测$\mathbf{y}_{\tilde{t}}$ 本质上是随机的，因为它们通常被视为来自某个隐藏空间和/或时间过程的一个样本；与此同时，观测中也包含随机误差，例如测量误差等。这意味着，一方面，$\mathbf{x}_t^A = \arg\min\ J = \mathfrak{G}(\mathbf{x}_t^B, \mathbf{y}_{\tilde{t}})$ 可以进一步简化为$\mathbf{x}_t^A = \mathcal{G}(\mathbf{y}_{\tilde{t}})$；另一方面，输入代价函数 J 或方程 (2) 的观测$\mathbf{y}_{\tilde{t}}$ 只是其样本空间 $\Omega_{\mathbf{y}_{\tilde{t}}}\ (\tilde{t} = t \sim (t+\Delta t))$ 中的样本元素 $\omega_{\mathbf{y}_{\tilde{t}}} \triangleq \mathbf{y}_{\tilde{t}}$，因此，我们可以用一种更一般的方式表示$\mathbf{x}_t^A = \mathcal{G}(\mathbf{y}_{\tilde{t}})$，以体现$\mathbf{y}_{\tilde{t}}$ 的随机性，如下所示。

$$\mathbf{X}_t^A = arg\ min\ J = \mathcal{G}(\mathbf{Y}_{\tilde{t}}) \tag{3}$$

几乎必然成立，其中大写字母表示随机变量，小写字母表示其对应的实现。根据方程（2），我们可以将方程（3）进一步展开为更明确的形式，即

$$\mathbf{X}_t^A = arg\ min\ J = \mathcal{G}(\mathbf{Y}_{\tilde{t}}) \propto C_t + \sum_{\tilde{t}=t}^{t+\Delta t} K_{\tilde{t}} \mathbf{Y}_{\tilde{t}}, \tag{4}$$

其中 C_t 表示一个常向量， $K_{\tilde{t}}$ 是由$\mathbf{HM}\ldots\mathbf{M}$组成的系数矩阵。通过这种方式，我们得到将被称为再分析输出的一般形式 $\mathbf{X}_t^A$的形式，其中 $t = 0, \Delta t, 2\Delta t, \ldots$。

因此，根据第 3.1 节引入的概念，我们知道，再分析输出 $\mathbf{x}_t^A = \mathbf{X}_t^A(\omega_{\mathbf{y}_{\tilde{t}}})$ 的每次运行都是来自一个平凡随机过程 $\mathbf{X}_t^A(t = 0, \Delta t, 2\Delta t, \ldots)$ 的一个实现。因此，总体而言，尽管我们不能直接将再分析输出的集合 $\{\mathbf{X}_0^A(\omega_{\mathbf{y}_{\tilde{t}(=0\sim\Delta t)}}), \mathbf{X}_{\Delta t}^A(\omega_{\mathbf{y}_{\tilde{t}(=\Delta t\sim 2\Delta t)}}), \mathbf{X}_{2\Delta t}^A(\omega_{\mathbf{y}_{\tilde{t}(=2\Delta t\sim 3\Delta t)}}) \ldots\}$ 视为一个随时间演化的随机过程的一个实现——这需要样本空间中的一个固定元素（例如 $\omega_{\mathbf{y}_{\tilde{t}(=0\sim\Delta t)}}$），而不是在所有按时间索引的随机变量$\mathbf{X}_t^A(t = 0, \Delta t, 2\Delta t, \ldots)$ 中沿时间索引变化的样本元素（即 $\omega_{\mathbf{y}_{\tilde{t}(=0\sim\Delta t)}}, \omega_{\mathbf{y}_{\tilde{t}(=\Delta t\sim 2\Delta t)}}, \ldots$）——但我们至少可以从每个平凡随机过程中获得不同实现的集合。

由此，我们或许能够进一步探讨这样一种可能性：将每个平凡随机过程各自不同实现所构成的集合，与时间随机过程的一个实现或样本路径等同起来；后者在时空随机建模领域更为常见。

如果我们只关注给定时间 t 的一个再分析输出，$\mathbf{x}_t^A \equiv \mathrm{X}_t^A(\omega_{\mathbf{y}_{\tilde{t}(=t\sim(t+\Delta t))}})$，根据第2节中四维变分资料同化的机制，我们知道，这个再分析输出是一个向量，收集了在一个可能无限但实际上有限的域$\{s_1, s_2, s_3, \ldots, s_N\} \subset \{s_1, s_2, s_3, \ldots\}$内所有空间网格位置上的大气场值。也就是说，在给定时间 t，我们有一个随空间演化的集合$\{\mathrm{X}_{s_1}^A(\omega_{\mathbf{y}_{\tilde{t}}};\ t), \mathrm{X}_{s_2}^A(\omega_{\mathbf{y}_{\tilde{t}}};\ t) \ldots \mathrm{X}_{s_N}^A(\omega_{\mathbf{y}_{\tilde{t}}};\ t)\}$，它是$\{\mathrm{X}_{s_1}^A(\omega_{\mathbf{y}_{\tilde{t}}};\ t), \mathrm{X}_{s_2}^A(\omega_{\mathbf{y}_{\tilde{t}}};\ t) \ldots\}$的一个子集合，$\tilde{t} = t \sim (t+\Delta t)$。在给定时间，沿不同空间索引演化的这样一个再分析输出集合，看起来与该给定时间的空间随机过程的一个实现非常相似 t。

4 随机性存在性的证明

迄今为止，我们的分析工作主要基于与随机过程相关的概念。为了严谨地证明四维变分资料同化的再分析输出（即欧洲中期天气预报中心哥白尼大气监测服务再分析数据）具有随机性，我们借助测度论。

在测度论中，随机过程可以通过两种途径构造：一种是通过定义在同一概率空间上的一组随机变量（或随机对象），另一种则是通过抽象动力系统（Gray 等，2009年，第 6–7 页）。现在，我们将从这两种视角分别证明，与我们的再分析输出相关的随机性确实存在。

4.1 视角 1：一系列随机变量

第 3.1 节指出，随机过程是随机对象（随机变量、随机向量或随机过程）的集合。但我们需要重新思考随机对象 究竟是什么，以及我们在第 3.2 节定义的再分析输出的一般形式 $\mathbf{X}_t^A(t = 0, \Delta t, 2\Delta t, \ldots)$ 是否确实符合随机对象这一概念。我们先阐述基本理论，随后再进行检验和证明。

根据 Gray 等人（2009年，第 4 页）、Billingsley（1995 年，第 182–186 页）以及 Athreya 等人（2006 年，第 192 页）的定义，给定两个可测空间 $(\Omega,\mathcal{A})$ 和 $(\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n))$，其中 Ω 表示样本空间，$\mathcal{A}$ 是 Ω 上的一个 σ-代数，$\mathbb{R}^n$ 是一个拓扑空间，而$\mathcal{B}(\mathbb{R}^n)$ 是 $\mathbb{R}^n$ 的子集所构成的一个博雷尔 σ-代数，若函数 $f : \Omega \to \mathbb{R}^n$ 是 (n $= 1$ 维) 随机向量，或其一维特例随机变量，则它是 $(\Omega,\mathcal{A})$ 上的一个可测函数，也就是说

$$f^{-1}(B) = \cap_{i=1}^{n}\{\omega : f_i(\omega) \in B\} \in \mathcal{A},\ \forall B \in \mathcal{B}(\mathbb{R}^n),$$

并且若 μ 是 $\mathcal{A}$ 上的一个测度，也就是说 $(\Omega,\mathcal{A}\ \mu)$ 是一个测度空间，那么存在 诱导测度 μf^{-1} ，定义在 $\mathcal{B}(\mathbb{R}^n)$ 上，使得

$$\mu f^{-1}(B) = \mu(f^{-1}(B)) = \mu(\cap_{i=1}^{n}\{\omega : f_i(\omega) \in B\}),\ \forall B \in \mathcal{B}(\mathbb{R}^n)$$

该测度由随机对象 f 诱导而来。这个诱导测度称为随机对象 f 的（联合）累积分布函数；它本身是单调不减且右连续的，因此是一个勒贝格–斯蒂尔杰斯测度。

并且，如果 $f : \mathbb{R}^i \to \mathbb{R}^k$ 是两个拓扑空间之间的连续映射，那么 f 是博雷尔可测的，参见 Billingsley（1995，第 183 页）的定理 3.2。

根据第 3.2 节的分析，我们知道$\mathbf{X}_t^A = \mathcal{G}(\mathbf{Y}\tilde{t}) \propto C_t + \sum_{\tilde{t}=t}^{t+\Delta t} K_{\tilde{t}}\mathbf{Y}\tilde{t}$，其中$\mathbf{Y}\tilde{t}$ 是一个由观测变量组成的向量，此处 n 等于网格位置的数量乘以大气成分的数量，$K_{\tilde{t}}$ 是一个由$\mathbf{HM}\ldots\mathbf{M}$ 构成的标量矩阵，其中涉及的$\mathbf{M}$ 项数取决于索引 $\tilde{t}$，而 C_t 是一个常数向量。详细推导见附录 C。因此，要验证再分析输出的一般形式$\mathbf{X}_t^A(t = 0,\ \Delta t,\ 2\Delta t,\ \ldots)$ 是否为随机对象，我们只需证明函数$\mathcal{G}(\cdot)$ 是可测的。

引理 1. 四维变分资料同化的再分析输出$\boldsymbol{X}_t^A(t = 0, \Delta t, 2\Delta t, \ldots)$ 的一般形式是一个随机对象。

证明. 我们首先记$\mathbf{Z}_{\tilde{t}} = K_{\tilde{t}}\mathbf{Y}_{\tilde{t}} = G_{\tilde{t}}(\mathbf{Y}\tilde{t})$，其中 $\tilde{t} \in [t,(t + \Delta t)]$，且 $G_{\tilde{t}}(\cdot)$ 是一个函数，因此 $\mathbf{X}_t^A = \mathcal{G}(\mathbf{Y}\tilde{t}) \propto C_t + \sum_{\tilde{t}=t}^{t+\Delta t} K_{\tilde{t}}\mathbf{Y}_{\tilde{t}} = C_t + \sum_{\tilde{t}=t}^{t+\Delta t} G_{\tilde{t}}(\mathbf{Y}\tilde{t})$。由于可测函数在加法、标量平移和标量乘法下封闭，因此，要证明$\mathbf{X}_t^A$ $=\mathcal{G}(\cdot)$可测，我们只需证明对于任意一个指标 $\tilde{t} \in [t, t + \Delta t]$，$\mathbf{Z}_{\tilde{t}} = G_{\tilde{t}}(Y_{\tilde{t}})$ 可测。˜不失一般性，我们取 $t = t$。

根据$\mathbf{Y}\tilde{t}$ 和 $\mathbf{Z}\tilde{t}$的定义，我们知道 $G_{\tilde{t}=t}(\cdot)$ 是两个拓扑空间之间的函数，即 $G_{\tilde{t}=t}$: $\mathbb{R}^n \to \mathbb{R}^n$，其中每个拓扑空间都配备了

由 n 开区间的笛卡尔积生成的 $Borel$ σ-代数 $\mathcal{B}(\mathbb{R}^n)$，特别是输出空间上的 σ-代数 $\mathcal{B}(\mathbb{R}^n)$ 为 $\mathcal{B}(\mathbb{R}^n) = \sigma\langle(-\infty, b_1) \times \ldots \times (-\infty, b_n)\rangle, (b_i\colon\ i = 1, \ldots n) \in \mathbb{R}^n$。又由于 $G_{\tilde{t}=t}(\mathbf{Y}_{\tilde{t}=t}) = K_t\mathbf{Y}_t$ 是线性映射，因此连续，故根据 Billingsley（1995，第183页）的定理3.2，它是 $Borel$ 可测的。这意味着

$$G_{\tilde{t}=t}^{-1}((-\infty, b_1) \times \ldots \times (-\infty, b_n)) = \cap_{i=1}^{n}\{\omega_{y_t} : [\ G_{\tilde{t}=t}(\omega_{y_t})]\ _i \in (-\infty, b_i)\}$$

是 博雷尔 可测的。因此 $\mathbf{X}_t^A = \mathcal{G}(\mathbf{Y}_{\tilde{t}}) \propto C_t + \sum_{\tilde{t}=t}^{t+\Delta t} K_{\tilde{t}}\mathbf{Y}_{\tilde{t}} = C_t + \sum_{\tilde{t}=t}^{t+\Delta t} G_{\tilde{t}}(Y_{\tilde{t}})$ 是 博雷尔可测的，因为每个 $G_{\tilde{t}}(Y_{\tilde{t}})$，$\tilde{t} \in [t, t + \Delta t]$ 都是博雷尔 可测的。$\mathbf{X}_t^A$是一个随机对象。

□

从上述证明中我们知道，无论 $\tilde{t}$ 仅为一个索引（例如 $\tilde{t} = t$），还是跨越一个区间（例如 $t \sim t + \Delta t$），都不会改变$\mathcal{G}$的可测性，因而也不会改变$\mathbf{X}_t^A$的随机性。因此，为了记号清晰和使用方便，我们只需为 $\tilde{t}$选择一个索引；不失一般性地，设 $\tilde{t} = t$，于是$\mathbf{X}_t^A = \mathcal{G}(\mathbf{Y}_{\tilde{t}=t}) = G_{\tilde{t}=t}(\mathbf{Y}_{\tilde{t}=t}) + C_t$，并将$\mathbf{Y}_t$ 的样本空间记为 Ω，其上配备$Borel$ σ-代数$\mathcal{A}$。$\mathbf{Y}_{\tilde{t}=t}$ 的一个样本元素 $\omega_{\mathbf{y}_{\tilde{t}=t}}$ 为方便起见记作 ω_t，因此在下文中，我们处理的是

$$\mathbf{X}_t^A = \mathcal{G}(\mathbf{Y}_t),$$

其中 $\mathcal{G} : (\Omega, \mathcal{A}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$，对应的 $\mathbf{X}_t^A$实现为 $\mathbf{x}_t^A = \mathbf{X}_t^A(\omega_t) = \mathcal{G}(\omega_t)$。

而且，如果 $(\Omega, \mathcal{A})$ 上存在概率测度 P，那么在 $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ 上必然存在由可测函数 $\mathcal{G}$诱导的概率测度 $P\mathcal{G}^{-1}$。

引理 2。 四维变分资料同化所产生的一般形式的再分析输出集合 $\{\boldsymbol{X}_t^A\colon\ t = 0, \Delta t, 2\Delta t, \ldots\}$ 是一个单边时间随机过程。

证明。 根据引理 1，每个 $\mathbf{X}_t^A$ 都是定义在共同概率空间 $(\Omega, \mathcal{A}\ P)$ 上的随机对象，且索引 t 从 0 开始跨越不同的正时间步。 □

这种时间随机过程本质上是现实世界中隐藏时间随机过程的数字化版本。根据关系 $\mathbf{x}_t^A = \mathbf{X}_t^A(\omega_t)$，我们知道，每次运行所得到的再分析输出集合 $\{\mathbf{x}_0^A, \mathbf{x}_{\Delta t}^A, \mathbf{x}_{2\Delta t}^A, \ldots\} = \{\mathbf{X}_0^A(\omega_0), \mathbf{X}_{\Delta t}^A(\omega_{\Delta t}), \mathbf{X}_{2\Delta t}^A(\omega_{2\Delta t}) \ldots\}$ 是来自每个平凡随机过程或每个单独随机变量的单个实现所组成的集合。在第 4.2.1 节的定理 2 中，我们将揭示它等价于时间随机过程的一个实现或样本路径。

此外，在给定时间 t，根据第2节中介绍的$\mathbf{x}_t^A$（或$\mathbf{x}_{k\Delta t}$）的定义——它是一个收集所有网格位置上的大气场列的所有值的向量——我们可以进一步根据空间索引集S，将给定时间 t 的随机对象$\mathbf{X}_t^A$展开为一个“更深层次”的集合，即$\{\mathbf{X}_t^A{}_{,s_1}, \mathbf{X}_t^A{}_{,s_2}, \mathbf{X}_t^A{}_{,s_3}, \ldots\}$，其中，这里的空间索引集$\{s_1, s_2, \ldots\}$在理论上是无限的，而在实际中是有限的，每个 $s_i \in \mathbb{R}^d$对应一个网格位置。由于$\mathbf{X}_t^A$是$\mathbf{Y}_t$的可测函数，因此很容易证明，$\mathbf{X}_t^A{}_{,s_i}$ $(i = 1, 2, \ldots)$同样是$\mathbf{Y}_t$的可测函数；因此，在给定时间 t，这些带有演化空间索引的随机向量集合$\{\mathbf{X}_{s_i}^A(t),\ i = 1, 2, \ldots\}$是一个d维空间离散随机场，或者简单地说，是一个空间随机场或空间随机过程（Shalizi，2007年，第4页）。每种大气成分对应一个。这里，将时间索引 t 移入括号中，是为了清晰地观察空间索引的演化。

引理 3. 在给定时刻 t，四维变分资料同化的具有一般形式的再分析输出 $\{\boldsymbol{X}_{s_i}^A(t), i = 1, 2, \ldots\}$ 是一个空间随机过程。

证明。 根据引理 1 和随机过程的定义。详见附录 D.1。 □

这种空间随机过程本质上是现实世界中隐藏空间过程的数字化版本。当选定有限个空间索引集 $\{s_1, s_2, \ldots, s_N\} \subset \{s_1, s_2, \ldots\} \subset \mathcal{S}$时，$\{\mathbf{X}^A_{s_1}(t), \mathbf{X}^A_{s_2}(t) \ldots, \mathbf{X}^A_{s_N}(t)\}$ 就是该空间随机过程的一个子集。

请注意，在给定时刻 t，样本元素 $\omega_t \in \Omega$ 随即固定。根据第 3.1 节中对随机过程的一个实现 的定义，我们知道，此时的再分析输出$\{\mathbf{x}^A_{s_1}(t)$，$\mathbf{x}^A_{s_2}(t)$，$\mathbf{x}^A_{s_3}(t) \ldots\}$ $\equiv \{\mathbf{X}^A_{s_1}(\omega_t;\ t)$，$\mathbf{X}^A_{s_2}(\omega_t;\ t)$，$\mathbf{X}^A_{s_3}(\omega_t;\ t) \ldots\}$ 实际上就是该空间随机过程的一个实现。

因此，我们得出了一个将再分析输出与空间随机过程的实现联系起来的重要结果，并将其总结如下，作为定理 1：

定理 1. 在给定时刻 t，通过 *4D-Var* 数据同化（即欧洲中期天气预报中心哥白尼大气监测服务再分析数据）得到的再分析输出 $\boldsymbol{x}^A_t = \{\boldsymbol{x}^A_{s_1}(t)$，$\boldsymbol{x}^A_{s_2}(t)$，$\boldsymbol{x}^A_{s_3}(t) \ldots\}$ 是空间随机过程 $\{\boldsymbol{X}^A_{s_1}(t)$，$\boldsymbol{X}^A_{s_2}(t)$，$\boldsymbol{X}^A_{s_3}(t) \ldots\}$ 的一个实现。

证明。 根据引理 3 和随机过程的一个实现的定义。详情见附录 D.2。

□

4.2 视角二：抽象动力系统

4.2.1 时间随机过程

根据 Gray 等人（2009年，第 7 页），一个抽象动力系统 由两个要素组成：其一是测度空间 $(\Omega, \mathcal{A}\ P)$，其二是在该空间上定义的可测可逆变换函数 T，形式为 $T\colon \Omega \to \Omega$。四元组 $(\Omega, \mathcal{A}\ P, T)$ 共同构成一个抽象动力系统，它反映了在测度空间 $(\Omega, \mathcal{A}\ P)$ 上反复应用可测变换 T 所产生的长期动态行为。由于可测函数的复合仍然是可测的，定义为 $T^r\omega = T(T^{r-1}\omega) = T \circ T^{r-1}(\omega)$ 的 T^r 相对于可测空间 $(\Omega,\mathcal{A})$ 也是可测的。

为将 抽象动力系统与随机过程联系起来，我们还需要一个 V值的[2]可测函数或随机变量 f，并假设它定义在同一个可测空间 $(\Omega, \mathcal{A})$ 上，那么定义为 $fT^r(\omega) = f(T^r\omega)$ 的新可测函数 $fT^r\colon \Omega \to V$，$\omega \in \Omega$，对于所有 $r \in \mathbb{Z}^+$ 显然都是一个随机变量。

这意味着一个 抽象动力系统 $(\Omega, \mathcal{A}\ P, T)$ 和一个可测函数 f 共同定义了一个单侧随机过程 $\{X_r\colon r \in \mathbb{Z}^+\}$，其中每个 $X_r(\omega) = fT^r(\omega) = f(T^r\omega)$，也就是说，$\mathrm{r}^{th}$ 实现值 of thisrandom process（即 $X_r(\omega)$）是可测函数 f 在原始样本空间 Ω 中经过 r-单位变换后的点（即 $f(T^r\omega)$）处的取值。对于我们而言，这将是一项关键性质，使我们能够将时间随机过程中的一个实现等同于来自每个单独随机变量或平凡随机过程的一组不同实现，从而得出我们在第 4.1 节中关于时间随机过程所得出的结论。

当 Ω 包含序列 $\{\omega_r\colon r \in \mathbb{Z}^+\}$ 时，T^l 会将序列 $\{\omega_r\colon r \in \mathbb{Z}^+\}$ 平移为序列 $\{\omega_{r+l}\colon r \in \mathbb{Z}^+\}$，其中每个坐标都向左平移 l 个时间单位。

并且，如果可测空间 $(\Omega, \mathcal{A})$ 是一个拓扑空间 $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$，那么 Krylov–Bogolyubov定理保证在$\mathcal{B}(\mathbb{R}^n)$上存在一个不变的博雷尔概率测-度 μ 使得 $\mu(T^{-1}(B)) = \mu(B)$，或者说，对于任意 $B \in \mathcal{B}(\mathbb{R}^n)$，都有 $T\mu = \mu$（Sinai 等，1989，第 4 页）。

现在将上述理论应用于我们的再分析输出，由此得出四维变分资料同化的再分析输出集合与时间随机过程的实现之间的关系，具体如下定理 2：

[2]当可测函数的输出空间未在上下文中明确说明时，我们将其记作一个取值于V 的随机变量。（Gray 等，2009年，第 2 页）

定理 2. 四维变分资料同化（即欧洲中期天气预报中心哥白尼大气监测服务的再分析数据）所得的再分析输出集合 $\{\boldsymbol{x}_0^A, \boldsymbol{x}_{\Delta t}^A, \boldsymbol{x}_{2\Delta t}^A, \ldots\}$ 确实是一个时间随机过程 $\{\boldsymbol{X}_t : \ t = 0, \Delta t, 2\Delta t, \ldots\}$ 的一个实现，也就是一条样本路径。

证明。从上一节第 4.1 节可知，再分析输出的一般形式为 $\mathrm{X}_t^A = \mathrm{G}(\mathrm{Y}_t)$，其中 G 是一个从 (Ω,A) 映射到 $(\mathbb{R}^n, \mathrm{B}(\mathbb{R}^n))$ 的 Borel 可测函数；根据 Y_t 和 X_t^A 的定义，此处 $(\Omega, \mathrm{A}) = (\mathbb{R}^n, \mathrm{B}(\mathbb{R}^n))$。令 $T^{\Delta t}$ 将序列的坐标向左移动 Δt 个单位，即 $T^{\Delta t}(\omega_0, \omega_{\Delta t}, \omega_{2\Delta t}, \ldots) = (\omega_{\Delta t}, \omega_{2\Delta t}, \omega_{3\Delta t} \ldots)$，$T^{2\Delta t}(\omega_0, \omega_{\Delta t}, \omega_{2\Delta t}, \ldots) = T^{\Delta t}(T^{\Delta t}(\omega_{\Delta t}, \omega_{2\Delta t}, \omega_{3\Delta t} \ldots)) = (\omega_{2\Delta t}, \omega_{3\Delta t} \ldots))$，一个常见的例子是 $\Delta t = 1$。

对于第一个时间步，即 $t = 0$，我们有 $\mathbf{x}_0^A = \mathbf{X}_0^A(\omega_0)$；对于第二个时间步$t = \Delta t$，我们有

$$\mathbf{x}_{\Delta t}^A = \mathbf{X}_{\Delta t}^A(\omega_{\Delta t}) \text{（根据随机变量（r.d.v.）的一个实现的定义）} = \mathcal{G}(\omega_{\Delta t}) \text{（根据} \mathbf{X}_{\Delta t}^A \text{ 的定义，且} \mathcal{G} \text{ 是一个可测函数）} = \mathcal{G}(T^{\Delta t}\omega_0) \text{（根据 } T, T^{\Delta t}\omega_0 = \omega_{\Delta t} \text{的定义）} = \mathcal{G}T^{\Delta t}(\omega_0) \text{（根据乘法的结合律）} = \mathbf{X}_{\Delta t}^A(\omega_0) \text{（通过动力系统和可测函数} \mathcal{G} \text{构造一个 r.d.v.）} \tag{5}$$

同样地，我们有

$$\mathbf{x}_{2\Delta t}^A = \mathbf{X}_{2\Delta t}^A(\omega_{2\Delta t}) = \mathcal{G}(\omega_{2\Delta t}) = \mathcal{G}(T^{2\Delta t}\omega_0) = \mathcal{G}T^{2\Delta t}(\omega_0) = \mathbf{X}_{2\Delta t}^A(\omega_0), \tag{6}$$

等等。

我们知道，动力系统 $(\Omega, \mathrm{A}\ P, T)$ 与可测函数 G 一起定义了一个随机变量，因此集合 $\{\mathrm{X}_0^A, \mathrm{X}_{\Delta t}^A, \mathrm{X}_{2\Delta t}^A, \ldots\}$ 是一个单边时间随机过程；根据第 3.1节中介绍的随机过程的一个实现或样本路径的定义，我们知道，对于给定的固定 ω_0，方程(5)、(6) 右侧的集合，即$\{\mathrm{X}_0^A(\omega_0), \{\mathrm{X}_{\Delta t}^A(\omega_0), \mathrm{X}_{2\Delta t}^A(\omega_0) \ldots\}$，确实是这个单边时间随机过程的一个实现或样本路径。另一方面，方程 (5)、(6) 左侧的集合是我们的再分析输出数据$\{\mathbf{x}_0^A, \mathbf{x}_{\Delta t}^A, \mathbf{x}_{2\Delta t}^A, \ldots\}$。因此，我们从四维变分资料同化获得的再分析输出集合确实是这个单边时间随机过程的一个实现。

□

因此，我们进一步完善了第 4.1 节中关于时间随机过程得出的结论。

4.2.2 空间随机过程

为了证明，在给定时刻 t ，四维变分资料同化的再分析输出是一个空间随机过程的实现，我们将讨论两种情形：一种是网格位置像一维中的时间索引一样完全有序，另一种是网格位置无序。

对于第一种情形，证明思路与定理 2 的证明类似，详见附录 D.3；对于第二种情形，则需要引入柱集等更多概念。为方便起见，以下所有证明均针对给定时刻 t的再分析输出，因此省略时间索引 t。

下面陈述网格位置无序时的证明思路。

令$\{s_{j'} : \ j' \in \mathcal{N}(s_j)\ j, \ j' \in \mathbb{Z}^+\}$ 表示位于位置 s_j 一阶邻域内的位置索引集合，例如，对于 $j = 1$，我们有$\{s_{j'} : \ s_{j'} \in \mathcal{N}(s_1)\}$，等等。随后，我们将它们收集到 $\mathcal{S} = \{\{s_{\mathcal{N}(s_1)}\}, \{s_{\mathcal{N}(s_2)}\} \ldots \{s_{\mathcal{N}(s_N)}\}\}$。

遵循 Athreya等人（2006 年，第 202 页）的定义，我们将一组实值函数 $\mathbb{R}^{\mathcal{S}}$ 定义为

$$\mathbb{R}^{\mathcal{S}} = \{\omega \mid \omega : s \to \mathbb{R}^{P}\},$$

此处， P 表示位置 s 在时间 t的气溶胶组分数量。此外，我们将 有限维柱集 *(f.d.c.s.)* $\mathcal{C}$ 定义为：对于 $\mathcal{S}_1 = \{s_1, s_2, \ldots, s_N\} \subset \mathcal{S}$， $\mathcal{C} \subset \mathbb{R}^{S}$，并且

$$\mathcal{C} = \{\omega : \omega \in \mathbb{R}^{S} \;\&\; (\omega(s_1), \ldots, \omega(s_N)) \in B\},$$

其中 B 是一个博雷尔集， $B \in \mathcal{B}(\mathbb{R}^{NP})$，正如我们在第 4.1 节中所见， $NP = n$。根据 Athreya 等人 (2006, p. 203)，我们知道，这类有限维柱集$\mathcal{C}$的集合构成一个 σ-代数 $\mathcal{R}^{S}$。对于一个 f.d.c.s.，根据 Carathéodory 扩张定理，$(\mathbb{R}^{S}, \mathcal{R}^{S}, P(\mathcal{C}))$ 是一个概率空间，其中 $P(\mathcal{C}) = \mu_{s_1,\ldots,s_N}(B)$。

然后我们定义投影映射 $\pi_{(s_1,s_2,\ldots,s_N)}$： $\mathbb{R}^{S} \to \mathbb{R}^{NP}$, $\forall (s_1, s_2, \ldots, s_N) \in \mathcal{S}^{N}$, $1 \leq N < \infty$,

$$\pi_{(s_1,s_2,\ldots,s_N)}(\omega) = (\omega(s_1), \omega(s_2), \ldots, \omega(s_N)),$$

特别是，对于一个位置 s, $\pi_s(\omega) = \omega(s) \triangleq \omega_s \equiv \mathbf{y}_s$，它是在位置$s$ 、时间 t的观测值，因此，集合 $\{\omega_s : s \in \mathcal{S}\} \subset \mathcal{A}$ 是 Ω 的 σ-代数，我们的可测函数 $\mathcal{G}$ 定义于其上，详见第 4.1 节。

我们还定义 Δs 为同一一阶邻域内任意两个位置 s_j,$s_{j'}$ 之间的欧几里得距离单位，即 $\Delta s = \parallel s_j - s_{j'} \parallel$；那么，可测可逆变换函数 $T : \Omega \to \Omega$ 可以将s_j一阶邻域内任意一个位置的 ω 平移 Δs个单位，并使其与 ω_{s_j}等价，即 $T^{\Delta s}\omega_{[\mathcal{N}(s_j)]_i} = \omega_{s_j}$。此处，$[\mathcal{N}(s_j)]_i$ 表示s_j一阶邻域内的任意一个位置。

因此，给定时间 t （省略）在某个位置s_j 的再分析输出为

$$\mathbf{x}^{A}_{s_j} = \mathbf{X}^{A}_{s_j}(\omega_{s_j}) = \mathcal{G}(T^{\Delta s}\omega_{[\mathcal{N}(s_j)]_i}) = \mathbf{X}_{s_j}(\omega_{[\mathcal{N}(s_j)]_i})$$

为便于说明，假设$\{s_2, s_3, s_4, s_5\}$ 位于 s_1的一阶邻域内，这也意味着 s_1 位于$\{s_2, s_3, s_4, s_5\}$中每个位置的一阶邻域内，于是有

$$\begin{aligned}
\mathbf{x}^{A}_{s_2} = \mathbf{X}^{A}_{s_2}(\omega_{s_2}) = \mathcal{G}(\omega_{s_2}) = \mathcal{G}(T^{\Delta s}(\omega_{[\mathcal{N}(s_2)]_i})) = \mathcal{G}T^{\Delta s}((\omega_{[\mathcal{N}(s_2)]_i})) = \mathbf{X}_{s_2}(\omega_{s_1}) \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
\mathbf{x}^{A}_{s_5} = \mathbf{X}^{A}_{s_5}(\omega_{s_5}) = \mathcal{G}(\omega_{s_5}) = \mathcal{G}(T^{\Delta s}(\omega_{[\mathcal{N}(s_5)]_i})) = \mathcal{G}T^{\Delta s}((\omega_{[\mathcal{N}(s_5)]_i})) = \mathbf{X}_{s_5}(\omega_{s_1})
\end{aligned}$$

从上述方程的最左侧，我们得到给定时间下、s_1的一阶邻域内所有位置上的再分析输出$\{\mathbf{x}^{A}_{s_2}, \mathbf{x}^{A}_{s_3}, \ldots \mathbf{x}^{A}_{s_5}\}$ 的集合；从最右侧，我们观察到固定样本元素 ω_{s_1}对应的空间演化随机变量的一个实现$\{\mathbf{X}_{s_2}(\omega_{s_1}), \mathbf{X}_{s_3}(\omega_{s_1}) \ldots, \mathbf{X}_{s_5}(\omega_{s_1})\}$。因此，给定时间 t 的四维变分资料同化再分析输出集合，是一个空间随机过程的实现。

为完成对四维变分再分析输出数据随机性的研究，我们将在下一节考察与每次再分析输出运行相关的不同误差及其相应特性。

5 误差及其特性

在第 2 节中，我们详细分析了再分析输出的每次运行是如何生成的。现在，我们以此为基础，进一步剖析相应误差之间的关联、分类、传播和相关机制。

5.1 按时间剖析误差

在 $t=0$、 $\tilde{t}=0\sim\Delta t$的第一次数据同化运行中，背景状态 $\mathbf{x}_0^B$、初始分析状态 $\mathbf{x}_0$和观测$\mathbf{y}^{\tilde{t}}$ 都有各自的误差来源。例如，背景状态$\mathbf{x}_0^B$和初始分析状态$\mathbf{x}_0$ 都被设为等于一个猜测值，而该猜测值本身具有固有的随机误差；此外，由于测量仪器和仪器类型等方面的差异，观测$\mathbf{y}^{\tilde{t}}$ 在˜期间收集的 $t=0\sim\Delta t$ 自然包含随机的观测测量误差。

与此同时，模拟观测 $\mathbf{H}_{\tilde{t}}^{o}\mathbf{x}^{\tilde{t}}$ 通过数值天气预报模型从 $\mathbf{x}^{\tilde{t}}$ 继承误差，即 $\mathbf{x}^{\tilde{t}}=\mathbf{M}^{\tilde{t}}\cdot\mathbf{M}^{\tilde{t}-\delta t}\cdots\mathbf{M}_{0+\delta t}\cdot\mathbf{x}_0,$，其中模型输入 $\mathbf{x}_0$ 存在固有随机误差，而模型本身也存在系统性偏差，这些偏差源于空间邻近性（来自$\mathbf{H}^o$）和时间邻近性（来自 $\mathbf{M}\ldots\mathbf{M}$），或数值天气预报模型模拟的时空分辨率有限。

而且，由于这些与误差相关的值 $\mathbf{x}_0^B$、$\mathbf{x}_0,\mathbf{y}^{\tilde{t}}$和 $\mathbf{H}_{\tilde{t}}^{o}\mathbf{M}^{\tilde{t}}\cdot\mathbf{M}^{\tilde{t}-\delta t}\cdots\mathbf{M}_{0+\delta t}\cdot\mathbf{x}_0$都参与了首个再分析输出 $\mathbf{x}_0^A$的生成，因此，这些相关误差（无论是随机误差还是系统性误差）也就成为首轮再分析输出$\mathbf{x}_0^A$中不可分割的一部分。

在第二次数据同化运行中，当 $t=\Delta t,\tilde{t}=\Delta t\sim 2\Delta t$ 时，背景状态 $\mathbf{x}_{\Delta t}^B$ 是将数值天气预报模型应用于第一次运行的再分析输出 $\mathbf{x}_0^A$ 而得到的，即 $\mathbf{x}_{\Delta t}^B = M_{\Delta t}\cdot M_{\Delta t-\delta t}\cdots M_{0+\delta t}\cdot\mathbf{x}_0^A$。因此，第一次运行的再分析输出 $\mathbf{x}_0^A$ 中包含的误差现在被传播到了第二次数据同化运行的背景状态中，即 $\mathbf{x}_{\Delta t}^B$。与上面对第一次运行的分析类似，这一误差传播过程既涉及数值天气预报模型中的偏差，也涉及模型输入 $\mathbf{x}_0^A$ 中的误差，而该模型输入来自数据同化的上一次运行。新收集的观测$\mathbf{y}^{\tilde{t}}$ 无疑也具有其自身的随机测量误差。所有这些误差再次成为第二次运行的再分析输出$\mathbf{x}_{\Delta t}^A$ 的一部分，使其无法摆脱这些误差。

相同的分析逻辑适用于数据同化的以下所有运行。我们在图1中总结并可视化了这些误差剖析及其传播。

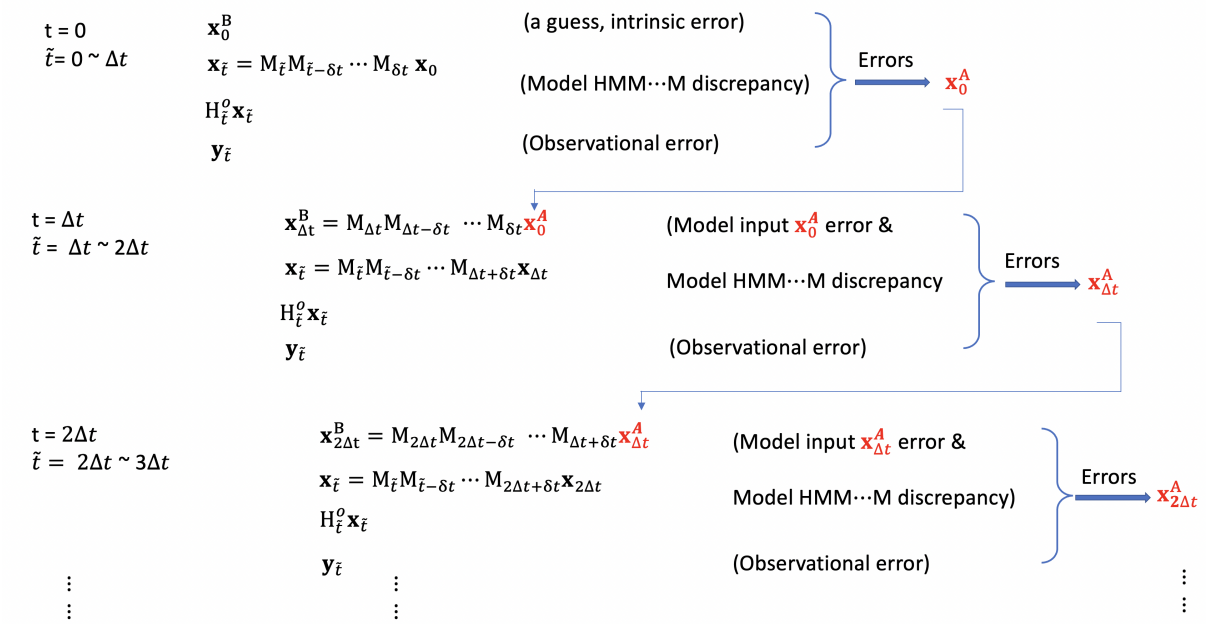


图1：与每次数据同化运行的再分析输出相关的误差分解与传播示意图。

因此，每次数据同化运行的再分析输出，或每个时间点的再分析输出$\{\mathbf{x}_0^A,\mathbf{x}_{\Delta t}^A,\mathbf{x}_{2\Delta t}^A,\ldots\}$ 都包含大量误差，主要包括数值天气预报模型输入误差、数值天气预报模型偏差和观测测量误差。初始状态中的随机误差

输入$\mathbf{x}_0$ 可通过利用一些监测观测进行校准，在一定程度上予以减少；参见 Rougier（2007年，第4节）和 Buizza（2000年，第3节），而不是纯粹依靠猜测，但模型偏差和观测测量误差则无法消除。

尽管每次再分析输出中的误差会传递到后续各次模型输入中，因此会在不同的数据同化运行之间或时间上相互关联，但这些误差的来源各不相同，因而可以合理地认为它们彼此独立。具体而言，对于首次初始化运行，数值天气预报模型输入$\mathbf{x}_0$ 误差源于建模人员的估计；模型偏差主要是由模型模拟在时空上的系统性邻近关系以及有限的时空分辨率造成的；观测测量误差则主要来自仪器等。因此，对于首次初始化运行的再分析输出，可以合理地假设数值天气预报模型输入误差、数值天气预报模型偏差和观测测量误差彼此独立，换言之，它们之间没有相互作用。从数学上说，对于首次再分析输出运行，

NWP 模型输入误差 ⫫ 数值天气预报模型偏差 ⫫ 观测测量误差 r,

其中，“⫫”表示“独立”。

然而，如果我们继续思考后续每一次数据同化运行，就会得到不同的情况。我们知道，本次运行的模型输入中的误差实际上 就是上一次运行的再分析输出中的误差，而它还可以进一步分解为上一次运行的模型输入误差、模型偏差和观测测量误差。我们还知道，一旦显式数值天气预报模型公式确定下来，不同数据同化运行之间的模型偏差就保持不变（见附录中公式（9）下方的讨论）。这些事实共同表明，本次运行的模型输入误差与本次运行的模型偏差相关，而后者与上一次运行的模型偏差相同。此外，由于观测测量误差和模型输入在每次运行中都会发生变化，因此仍可假定本次运行的模型输入误差与本次运行的观测测量误差相互独立；与此同时，由于二者来源相互独立，本次运行的模型偏差与观测测量误差也仍然相互独立。用数学形式表示，对于除第一次初始运行之外的每一次后续数据同化运行，

数值天气预报模型输入误差 ⫫ 数值天气预报模型偏差 数值天气预报模型输入误差 ⫫ 观测测量误差 数值天气预报模型偏差 ⫫ 观测测量误差

5.2 误差的空间分解

为分析空间上的误差，我们选择某一时刻 t。

此时，再分析输出为$\mathbf{x}_t^A$，在有限空间域内可进一步展开为 $\mathbf{x}_t^A \equiv \{\mathbf{x}_{s_1}^A(t),\ \mathbf{x}_{s_2}^A(t) \ldots\ldots \mathbf{x}_{s_N}^A(t)\}$。如上文分析，<style id='x'>$\mathbf{x}_t^A$从$\mathbf{x}_t^B$、$\mathbf{x}_t$、$\mathbf{y}^{\tilde{t}}$以及$\mathbf{H}_{\tilde{t}}^{o}\cdot\mathbf{M}_{\tilde{t}}\cdot\mathbf{M}_{\tilde{t}-\delta t}\cdots\mathbf{M}_{t+\delta t}\cdot\mathbf{x}_t$继承的误差，现在分布在每个网格位置的各个再分析值上。从上文第 5.1 节可知，在每个网格位置，与再分析输出相关的误差主要可分为数值天气预报模型输入误差、数值天气预报模型偏差和观测测量误差；由于对应的误差来源不同，这三类误差各自具有特定的依赖/独立关系。

此外，如果我们将一次再分析输出中包含的这三类误差作为一个整体来看，那么这些误差在不同网格位置之间的相关性有两种特殊情况：

1. 不同网格的再分析输出都具有相同数量的误差，因此不同网格之间再分析输出误差的相关性都等于一；

2.不同网格上的再分析输出都各自具有不同的个体误差，因此不同网格间再分析输出误差的相关性均等于零，这等同于按网格独立的误差。

一般情况下，与某次再分析输出相关的误差在不同网格位置之间的相关性介于 0 和 1 之间。而在每个网格上，与该特定网格和特定时刻的再分析输出相关的误差，仍可根据第 5.1 节中的分析，按照不同来源归为不同类型。

5.3 按组成对误差进行剖析

如第 2 节所述，在给定时刻 t 和给定网格位置 s，$\mathbf{x}_s^A(t) = \mathbf{X}_s^A(\omega_{\mathbf{y}_{\bar{t}}}; t)$ 汇集了不同大气成分的所有数值，即 $\mathbf{x}_s^A(t)\,[X_s^{A_{PM25}}\ \omega_{\mathbf{y}_{\bar{t}}}\ t\quad X_s^{A_{BC}}\ \omega_{\mathbf{y}_{\bar{t}}}\ t\quad X_s^{A_{SU}}\ \omega_{\mathbf{y}_{\bar{t}}}\ t\quad \ldots] = [x_s^{A_{PM25}}\ t\quad x_s^{A_{BC}}\ t\quad x_s^{A_{SU}}\ t\quad \ldots]$

$\mathbf{X}_s^A(\,;\,t)\,(\,;\,)\,(\,;\,)\,(\,;\,)\,(\,)\,(\,)\,(\,)$，其中这些标量各自都具有空间特征和时间特征。

因此，在给定时刻 t 和网格 s 位置上，与再分析输出$\mathbf{x}_s^A(t)$相关的误差会传递到该时刻、该网格位置上各大气成分的值中。两种极端情况是：要么该时刻该网格位置上的每种大气成分都具有相同大小的误差，也就是说，此处此时不同大气成分之间的误差相关性均为 1；要么相关性均为 0，这意味着此处此时不同大气成分之间的再分析输出误差完全不同。一般而言，给定位置、给定时刻不同大气成分之间的误差相关性介于 0 和 1 之间。

与上述情况相同，给定网格、给定时刻每种大气成分相关的误差也可根据不同误差来源分为不同类型，而且不同类型的误差之间存在相互依赖或相互独立的关系。

5.4 误差协方差矩阵在 J 中的作用

请注意，第 2 节提到的代价函数 J 中两个误差协方差矩阵$\mathbf{B}$ 和$\mathbf{R}$ 的作用，仅仅是对背景状态 $\mathbf{x}^B$ 和观测值$\mathbf{y}$ 中的误差进行重新缩放，使其标准化以便进行一般性比较；但是，乘以这两个矩阵的逆矩阵（即$\mathbf{B}^{-1}$ 和$\mathbf{R}^{-1}$）无法消除$\mathbf{x}^B$ 和$\mathbf{y}$中的随机误差。

6 讨论

本研究的动机在于，欧洲中期天气预报中心哥白尼大气监测服务再分析数据集的随机性尚不明确，而该数据集包含丰富的信息，可供许多环境相关领域的研究人员使用，例如时空建模、公共卫生、气候变化、环境智能等。本论文旨在研究通过四维变分资料同化机制由计算机生成的这些再分析输出的随机性。

我们首先清晰阐述了四维变分资料同化机制，后续的证明和推理均以此为基础。

我们采用测度论，从两个角度证明了这些再分析输出中随机性的存在。具体而言，从随机变量序列的角度出发，我们证明了与四维变分资料同化的再分析输出相关的时空随机过程确实存在；这些过程本质上是现实世界中隐含的时空过程的数字化版本，见引理 1、2 和 3。特别是，我们确认在给定时刻 t，相应的再分析输出是某个空间随机过程的一次实现，见定理 1；从抽象动力系统的角度出发，我们进一步揭示，

四维变分资料同化所有运行的再分析输出集合是某个时间随机过程的一次实现，见定理 2。从这一角度还证明了空间过程的存在，以及空间过程的一次实现，见第 4.2.2 节。

这些结果意味着，在实践中，我们可以将欧洲中期天气预报中心的哥白尼大气监测服务CAMS再分析数据集视为与从监测设备获得的观测值相同；这些观测值通常被认为具有内在随机性，因此，随机时空模型适用于这一欧洲中期天气预报中心的哥白尼大气监测服务CAMS再分析数据集。

我们还全面分析了与这些再分析输出相关的不同类型误差的不同来源，详见第 5节。总体而言，它们包括可约减的模型输入误差，以及不可约减的模型偏差和观测测量误差。此外，我们还阐明了这三类误差之间相互依赖或相互独立的关系，这些关系共同为误差项的建模提供了明确指导。具体而言，模型偏差与观测测量误差之间的相互独立性使我们能够使用加性误差项对再分析输出数据进行建模，而无需对它们之间的任何交互作用进行建模。

总而言之，既然已证实再分析数据集中存在空间和时间随机过程，且模型偏差与测量误差相互独立，那么这些标准的时空随机建模框架，例如数据 $=$ 协变量 $+$ 空间随机效应 $+$ 时间随机效应 $+$*(时空交互作用)* $+$ 随机测量误差（Cressie 等，2015，第 304–305 页），适用于欧洲中期天气预报中心哥白尼大气监测服务再分析数据集。

因此，我们将四维变分资料同化的再分析输出（即欧洲中期天气预报中心哥白尼大气监测服务再分析数据集）的实用范围，扩展到气候学计算（均值、百分位数等）、趋势研究、地理可视化制图等经验性用途，以及不确定性量化和数据融合等有限的统计应用之外。相反，这些再分- 析数据可以通过随机统计模型进行建模，尤其是单独使用随机时空模型，而无需融合或集合。

从第 5节的误差分析中，我们认识到，在再分析数据$\mathbf{x}_t^A$的生成过程中，空间和时间信息由两个矩阵 $\mathbf{H}$ 和$\mathbf{M}$ 参与处理，而该过程是非随机的。通过我们的证明，我们知道再分析输出$\mathbf{x}_t^A$本质上是现实世界中隐藏的空间和时间过程的数字化版本的实现，因此，使用随机建模方案实际上可以更好地处理这两类信息。

与此同时，在第 5.3 节中，我们还发现，在给定时间和网格上，不同气溶胶组分之间的误差通常存在相关性；然而，我们在第 2 节提到，出于实际应用上的便利，四维变分资料同化采用单变量方式实施。因此，多变量随机时空建模方案或许能进一步改善再分析输出数据，而改进后的结果又可惠及更广泛的环境影响研究，例如公共卫生、气候变化和环境智能等。

仅凭本文的结论，也足以为时空建模人员和环境人工智能研究者提供有力的理论依据，使他们在打算使用这一欧洲中期天气预报中心哥白尼大气监测服务再分析数据集进行随机模型研究时，能够直接着手开展工作。

需要指出的是，尽管本文强调的是四维变分资料同化机制，但用于生成欧洲中期天气预报中心哥白尼大气监测服务再分析数据集的实际方法，更准确地说是增量 四维变分。不过，增量四维变分的关键特征、函数结构和性质都与四维变分完全相同。增量四维变分只是对代价函数 J 的状态分量（即 $\mathbf{x}_t^B$、$\mathbf{x}_t$ 和 $H_t^o\mathbf{x}_t$）做了一些细微修改：将每个分量拆分为自行定义的参考状态 $\mathbf{x}_t^R$ 和增量状态 $\delta\mathbf{x}_t$，并将最小化目标从$\mathbf{x}_t$ 改为 $\delta\mathbf{x}_t$；同时，将数值天气预报模型$\mathbf{M}$ 设为线性形式，而非四维变分中的非线性形式。因此，将四维变分改为增量形式的主要目的是减轻

计算负担。由于原始四维变分的各个分量和结构含义更易于理解（见第 2 节），因此示范效果也更清晰，我们主要着重介绍四维变分。有关增量四维变分的完整细节，请参见 Bannister（2022年，第 17–19 页）。

此外，正如 Wikle 等人（2007年，第 2.4 节）所指出的，优化问题可以等价地表示为变分问题；对于高维任务，变分方法在计算上更加高效。因此，在实际应用中，四维变分数据同化再分析采用变分推断方法来获得输出结果（Bannister，2001年，第 3.3 节）。

另需注意，尽管本文中的证明主要从实际同化运行的角度，关注在时间和空间上离散随机过程的存在性，但只需将时间增量 Δt 设为无限小，即可轻松将这些证明推广至连续情形，从而得到连续时间过程；对于连续空间过程，则可以通过增域渐近理论（Cressie, 1993年, p. 350）扩大空间域，或者通过填充渐近理论（Cressie, 1993年, p. 350），在有限域内不断增加网格位置的数量 N。相应的符号变化只需将下标（例如 X_t, X_{s_i}）替换为函数括号中的参数，即 $X(t)$, $X(s_i)$。

关于未来工作，实际上还有其他思路可以证明四维变分资料同化再分析数据中随机性的存在。

请注意，在第 3.2 节中，我们将背景状态 $\mathbf{x}_t^B$ 视为常数，这是因为它取自上一次运行的再分析输出，而其随机性尚未得到证明；我们只关注随机观测 $\mathbf{y}^{\tilde{t}}$。或者，我们也可以忽略每次运行中的同化观测 $\mathbf{y}^{\tilde{t}}$，仅关注时刻 t 的再分析输出 $\mathbf{x}_t^A$ 与时刻 t 的背景状态 $\mathbf{x}_t^B$ 之间的关系，而后者本质上就是时刻 $t - \Delta t$ 的再分析输出，这便是一个经典的确定性动态系统。然而，即使从这个确定性系统出发，只要该确定性系统的初始状态是随机的，我们仍然能够将其与随机性联系起来，例如参见 Berliner，1992年。随机初始状态通常可以通过我们对系统初始化时所知信息的不完备性来解释，并且通常会极大地影响确定性系统明确且唯一地确定其未来值的能力，因此通常会导致一种称为确定性混沌 的状态（Chan 等，2001年，第 1 页），在这种状态下，人们通常无法区分混沌随机性与随机过程随机性。Cressie 等（2015年，第 58–59 页）也讨论了这一点。

致谢

本研究由艾伦·图灵研究所通过图灵博士奖学金资助。第一作者感谢 Peter Ashwin 教授对证明提供的建议。第一作者也感谢 Antje Inness 博士提供了实际应用中有关时间演化矩阵的背景信息。

参考文献

Athreya, Krishna B 和 Soumendra N Lahiri（2006）。测度论与概率论。第19卷。Springer。Banerjee, Sudipto、Bradley P Carlin 和 Alan E Gelfand（2014）。空间数据的层次建模与分析。CRC出版社。Bannister, Ross（2022）。变分数据同化 *I*：背景与方法。讲义。网址：http://www.met.reading.ac.uk/~darc/nceo_training_2022/Var_RNB_Lecture1_v1.pdf （访问于2022年8月26日）。Bannister, Ross N（2001）。“基础四维变分”。载于：雷丁：雷丁大学。Berliner, L Mark（1992）。“统计学、概率与混沌”。载于：《统计科学》，第69–90页。Berrocal, Veronica J、Alan E Gelfand 和 David M Holland（2010）。“数值模型输出的时空降尺度模型”。载于：《农业、生物与环境统计学杂志》15.2，第176–197页。Billingsley, P.（1995）。《概率与测度》。《概率与统计学系列》。Wiley。ISBN：9780471007104。网址：https://books.google.co.uk/books?id=z39jQgAACAAJ。Buizza, Roberto（2000）。《混沌与天气预报》*2000*年*1*月。网址：https://www.ecmwf.int/sites/default/files/elibrary/2002/16927-chaos-and-weather-prediction.pdf （访问于2022年7月20日）。Chan,Kung-Sik 和 Howell Tong（2001）。《混沌：统计学视角》。Springer Science & Business Media。Courtier, Philippe、J-N Th´epaut 和 Anthony Hollingsworth（1994）。“采用增量方法实施四维变分的业务化策略”。载于：《英国皇家气象学会季刊》120.519，第1367–1387页。Cressie, Noel 和 Christopher K Wikle（2015）。《时空数据统计学》。John Wiley & Sons。Cressie, Noel AC（1993）。“空间数据统计学”。载于：《空间数据统计学》，第900–900页。欧洲中期天气预报中心（无日期）。哥白尼大气监测服务。网址：https://www.ecmwf.int/en/about/what-we-do/environmental-services/copernicus-atmosphere-monitoring-service （访问于2022年8月8日）。Gray, Robert M 和 RM Gray（2009）。《概率、随机过程与遍历性质》。第1卷。Springer。Grimmett, Geoffrey 和 David Stirzaker（2001）。“概率与随机过程”。载于。Inness, Antje 等（2019）。“CAMS大气成分再分析”。载于：《大气化学与物理》19.6，第3515–3556页。Kalnay, Eugenia（2003）。《大气建模、数据同化与可预报性》。剑桥大学出版社。Kennedy, Marc C 和 Anthony O’Hagan（2001）。“计算机模型的贝叶斯校准”。载于：《皇家统计学会杂志：*B*辑（统计方法学）》63.3，第425–464页。Lorenz, Edward N（1963年）。“确定性非周期流”。载于：《大气科学杂志》20.2，第130–141页。Rougier, Jonathan（2007）。“利用气候模型评估集合对未来气候进行概率推断”。载于：《气候变化》81.3，第247–264页。

Sain, Stephan R、Reinhard Furrer 和 Noel Cressie（2011）。“区域气候模型多变量输出的空间分析”。载于：应用统计学年刊，第 150–175 页。Shaddick, Gavin 等（2020）。“全球空气质量：一种用于疾病负担分析暴露评估的跨学科方法”。载于：大气12.1，第 48 页。Shalizi, Cosma（2007）。高级概率论 *II*。讲义。网址：https://www.stat.cmu.edu/~cshalizi/754/notes/all.pdf（访问于 18/08/2022）。Sinai, Yakov Grigor’evˇıc 和 Ya G Sinai（1989）。动力系统 *II*：遍历理论及其在动力系统与统计力学中的应用。Springer。Wackernagel, Hans（2013）。多元地统计学：导论及应用。Springer Science & Business Media。Wikle, Christopher K 和 L Mark Berliner（2007）。“数据同化贝叶斯教程”。载于：物理学 *D*：非线性现象230.1-2，第 1–16 页。Zidek, James V、Nhu D Le 和 Zhong Liu（2012）。“时空场数据与模拟数据的结合：臭氧应用”。载于：环境与生态统计学19.1，第 37–56 页。

附录

数值天气预报导致的模式状态随时间演变

第二个二次项，例如代价函数 J (1) 中的 $\sum_{\tilde{t}=t}^{t+\Delta t}(\mathbf{y}_{\tilde{t}}-H_{\tilde{t}}^{o}\mathbf{x}\tilde{t})^{T}(\mathbf{y}_{\tilde{t}}-H_{\tilde{t}}^{o}\mathbf{x}\tilde{t})$，不仅涉及时间 t的模式状态 $\mathbf{x}_t$ ，还涉及时间窗口内离散增量时间步 $t+\delta t$ 上的模式状态 $\mathbf{x}_{t+\delta t}$ ，例如 $t\sim($ $t+\Delta t)$。此类模式状态 $\mathbf{x}_{t+\delta t}$ 通过由物理定律驱动的函数形式 $m(.)$（因此通常是非线性的）从较早的状态$\mathbf{x}_t$ 演化获得，即对于任何时间步 t,

$$\mathbf{x}_{t+\delta t}=m(\mathbf{x}_t).$$

上述方程表明，向量$\mathbf{x}_{t+\delta t}$ 中的每个分量都是向量$\mathbf{x}_t$各分量的非线性组合。

随后，使用在某一点（例如$\mathbf{x}_t^B$或简单地取$\mathbf{0}$）展开的泰勒级数，以线性项近似非线性 $m(\mathbf{x}_t)$。

例如，通过在$\mathbf{x}_t^B$处对$\mathbf{x}_{t+\delta t}=m(\mathbf{x}_t)$进行泰勒展开，我们得到

$$\begin{aligned}\mathbf{x}_{t+\delta t}&=m(\mathbf{x}_t)\\&=m(\mathbf{x}_t^B)+m'(\mathbf{x}_t^B)(\mathbf{x}_t-\mathbf{x}_t^B)+o(\cdot),\end{aligned}$$

这里， $m'(\mathbf{x}_t^B)$是 $m(\mathbf{x}_t)$ 关于向量$\mathbf{x}_t$ 的每个分量的导数，并在 $\mathbf{x}_t=\mathbf{x}_t^B$处进行计算。作为示例，我们假设 $\mathbf{x}_t$ 是一个二维向量，那么

$$m'(\mathbf{x}_t^B)=\left[\begin{array}{cc}\frac{\partial[m(\mathbf{x}_t)]_1}{\partial[\mathbf{x}_t]_1} & \frac{\partial[m(\mathbf{x}_t)]_1}{\partial[\mathbf{x}_t]_2}\\ \frac{\partial[m(\mathbf{x}_t)]_2}{\partial[\mathbf{x}_t]_1} & \frac{\partial[m(\mathbf{x}_t)]_2}{\partial[\mathbf{x}_t]_2}\end{array}\right]\Bigg|_{\mathbf{x}_t=\mathbf{x}_t^B}\triangleq M$$

So,

$$\begin{aligned}\mathbf{x}_{t+\delta t}&=m(\mathbf{x}_t)\\&=m(\mathbf{x}_t^B)+M(\mathbf{x}_t-\mathbf{x}_t^B)+o(\cdot).\end{aligned}$$

如果在$\mathbf{0}$处展开泰勒级数，那么

$$\begin{aligned}\mathbf{x}_{t+\delta t}&=m(\mathbf{x}_t)\\&=m(\mathbf{0})+M(\mathbf{x}_t)+o(\cdot)\\&\approx M\mathbf{x}_t,\end{aligned}$$

其中，矩阵 M 包含在$\mathbf{0}$处对 $m(\mathbf{x}_t)$求得的所有已知一阶导数。

更具体地，将$\mathbf{x}_{t+\delta t}=M\mathbf{x}_t$ 中的 M 记为 $M_{t+\delta t}$ ，并按照用在$\mathbf{0}$处的泰勒展开近似一个模式状态的相同思路，可得

$$\begin{aligned}\mathbf{x}_{t+\delta t}&=M_{t+\delta t}\mathbf{x}_t\\&=M_{t+\delta t}M_t\mathbf{x}_{t-\delta t}\qquad(\text{by }\mathbf{x}_t=M_t\mathbf{x}_{t-\delta t})\\&=M_{t+\delta t}M_tM_{t-\delta t}\mathbf{x}_{t-2\delta t}\qquad(\text{by }\mathbf{x}_{t-\delta t}=M_{t-\delta t}\mathbf{x}_{t-2\delta t})\\&\cdots\\&=M_{t+\delta t}M_tM_{t-\delta t}\ldots M_{2\delta t}\mathbf{x}_{\delta t}\\&=M_{t+\delta t}M_tM_{t-\delta t}\ldots M_{2\delta t}M_{\delta t}\mathbf{x}_0\qquad(\text{by }\mathbf{x}_{\delta t}=M_{\delta t}\mathbf{x}_0)\end{aligned}$$

因此，模式状态 $\mathbf{x}\tilde{t}$ 随时间演变为 $\mathbf{x}_t=M_tM_{t-\delta t}\ldots M_{\delta t}\mathbf{x}_0$的描述。

请注意，在实际应用中， $m(\cdot)$通常会简化为线性形式。

B $\mathbf{H}_t^o$ 的结构

令$\mathbf{s}_i = (x_i, y_i)$为每个网格的质心。可以获取各个质心处的观测值，例如，在质心(x_2, y_i)、(x_3, y_i)处，我们分别得到观测值 O_{2i}、 O_{3i}。我们想知道不在质心处、而是在位置(x_T, y_i)的观测值 O_{Ti}，该位置位于(x_2, y_i)和(x_3, y_i)之间，即 $x_2 < x_T < x_3$。这需要使用插值系数对两个观测值 O_{2i}、O_{3i}进行平滑处理，也就是 $O_{Ti} = \alpha O_{2i} + \beta O_{3i}$。这里，$\alpha = \frac{x_3 - x_T}{x_3 - x_2}$，且 $\beta = \frac{x_T - x_2}{x_3 - x_2}$。这些插值系数是$\mathbf{H}_t^o$中的元素。

我们可以看到，如果假设这些位置已排序，则$\mathbf{H}_t^o$ 中的每个元素都位于 (0,1) 内。

C Il代价函数 J 和 $\nabla_{\mathbf{x}} J$的说明性推导

为便于清晰地检查，我们省略背景状态误差协方差矩阵$\mathbf{B}$ 和观测误差协方差矩阵$\mathbf{R}$ ，它们都只是标量；并暂且假设$\mathbf{x}$和$\mathbf{y}$ 都是一维向量，而非 n 维向量，以便更清楚地观察其二次结构。

在第一次运行中， $t = 0, \tilde{t} = 0 \sim \Delta t$,

$$
\begin{aligned}
J[x_0] &\propto \frac{1}{2}(x_0^B - x_0)^2 + \frac{1}{2}\{(y_0 - H_0^o x_0)^2 + (y_{\delta t} - H_{\delta t}^o x_{\delta t})^2 + \ldots + (y_{\Delta t} - H_{\Delta t}^o x_{\Delta t})^2\} \\
&= \frac{1}{2}(x_0^B - x_0)^2 + \frac{1}{2}\{(y_0 - H_0^o x_0)^2 + (y_{\delta t} - H_{\delta t}^o M_{\delta t} x_0)^2 + \ldots + (y_{\Delta t} - H_{\Delta t}^o M_{\Delta t} M_{\Delta t - \delta t} \ldots M_{\delta t} x_0)^2\}
\end{aligned}
$$

对 x_0 求一阶导数

$$
\begin{aligned}
\nabla_{x_0} J &\propto -(x_0^B - x_0) - H_0^o(y_0 - H_0^o x_0) - \ldots + (-H_{\Delta t}^0 M_{\Delta t} M_{\Delta t - \delta t} \ldots M_{\delta t})(y_{\Delta t} - H_{\Delta t}^0 M_{\Delta t} M_{\Delta t - \delta t} \ldots M_{\delta t} x_0) \\
&= -(x_0^B - x_0) - \sum_{\tilde{t}=0}^{\Delta t} H_{\tilde{t}}^o M_{\tilde{t}} M_{\tilde{t} - \delta t} \ldots M_{\delta t + 0}(y_{\tilde{t}} - H_{\tilde{t}}^o M_{\tilde{t}} M_{\tilde{t} - \delta t} \ldots M_{\delta t + 0} x_0),\ M_0 = 1.
\end{aligned}
$$

在第二次运行中， $t = \Delta t, \tilde{t} = \Delta t \sim 2\Delta t$,

$$
\begin{aligned}
J[x_{\Delta t}] &\propto \frac{1}{2}(x_{\Delta t}^B - x_{\Delta t})^2 + \frac{1}{2}\sum_{\tilde{t}=\Delta t}^{2\Delta t}(y_{\tilde{t}} - H_{\tilde{t}}^o x_{\tilde{t}})^2 \\
&= \frac{1}{2}(x_{\Delta t}^B - x_{\Delta t})^2 + \frac{1}{2}\{(y_{\Delta t} - H_{\Delta t}^o x_{\Delta t})^2 + (y_{\Delta t + \delta t} - H_{\Delta t + \delta t}^o x_{\Delta t + \delta t})^2 + \ldots + (y_{2\Delta t} - H_{2\Delta t}^o x_{2\Delta t})^2\} \\
&= \frac{1}{2}(x_{\Delta t}^B - x_{\Delta t})^2 + \frac{1}{2}\{(y_{\Delta t} - H_{\Delta t}^o x_{\Delta t})^2 + \ldots + (y_{2\Delta t} - H_{2\Delta t}^o M_{2\Delta t} M_{2\Delta t - \delta t} \ldots M_{\Delta t + \delta t} x_{\Delta t})^2\}
\end{aligned}
$$

对上述每一项分别求关于 $x_{\Delta t}$ 的一阶导数，并以紧凑的求和形式写出

$$
\nabla_{x_{\Delta t}} J \propto -(x_{\Delta t}^B - x_{\Delta t}) - \sum_{\tilde{t}=\Delta t}^{2\Delta t} H_{\tilde{t}}^o M_{\tilde{t}} M_{\tilde{t} - \delta t} \ldots M_{\Delta t + \delta t}(y_{\tilde{t}} - H_{\tilde{t}}^o M_{\tilde{t}} M_{\tilde{t} - \delta t} \ldots M_{\Delta t + \delta t} x_{\Delta t}),\ M_{\Delta t} = 1.
$$

对于在 n 维向量中关于时刻 t 的所需初始状态 $\mathbf{x}_t$ 的一阶导数的一般形式，我们有

$$
\nabla_{\mathbf{x}_t} J \propto -(\mathbf{x}_t^B - \mathbf{x}_t) - \sum_{\tilde{t}=t}^{t+\Delta t} (\mathbf{H}_{\tilde{t}}^o \mathbf{M}_{\tilde{t}} \mathbf{M}_{\tilde{t} - \delta t} \ldots \mathbf{M}_{t + \delta t})^T (\mathbf{y}_{\tilde{t}} - \mathbf{H}_{\tilde{t}}^o \mathbf{M}_{\tilde{t}} \mathbf{M}_{\tilde{t} - \delta t} \ldots \mathbf{M}_{t + \delta t} \mathbf{x}_t), \qquad (7)
$$

˜其中$\mathbf{M}_{\tilde{t}} = I$ 每当 $t = t_0$。

要获得所需的初始状态$\mathbf{x}_t$ 在每个时刻 t，将上述方程（7）设为 $\mathbf{0}$ 并重新整理，

$$\mathbf{x}_t^B + \sum_{\tilde{t}=t}^{t+\Delta t}(\mathbf{H}_{\tilde{t}}^o\mathbf{M}_{\tilde{t}}\mathbf{M}_{\tilde{t}-\delta t}\dots\mathbf{M}_{t+\delta t})^T\mathbf{y}_{\tilde{t}} = (I + \sum_{\tilde{t}=t}^{t+\Delta t}(\mathbf{H}_{\tilde{t}}^o\mathbf{M}_{\tilde{t}}\mathbf{M}_{\tilde{t}-\delta t}\dots\mathbf{M}_{t+\delta t})^T(\mathbf{H}_{\tilde{t}}^o\mathbf{M}_{\tilde{t}}\mathbf{M}_{\tilde{t}-\delta t}\dots\mathbf{M}_{t+\delta t}))\mathbf{x}_t,$$

因此，期望的初始状态$\mathbf{x}_t$ 之前的系数矩阵 $(I + \sum_{\tilde{t}=t}^{t+\Delta t}(\mathbf{HM}\dots\mathbf{M})^T(\mathbf{HM}\dots\mathbf{M})$ 必须是非奇异的，也就是说必须可逆，所以

$$\mathbf{x}_t \propto (I + \sum(\cdot)^T(\cdot))^{-1}\mathbf{x}_t^B + \sum_{\tilde{t}=t}^{t+\Delta t}(I + \sum(\cdot)^T(\cdot))^{-1}(\mathbf{H}_{\tilde{t}}^o\mathbf{M}_{\tilde{t}}\mathbf{M}_{\tilde{t}-\delta t}\dots\mathbf{M}_{t+\delta t})^T\mathbf{y}_{\tilde{t}} \tag{8}$$

其中，$(\cdot)^T(\cdot)$ 中的“·”是 $\mathbf{HM}\dots\mathbf{M}$。为方便起见，我们还可以将$\mathbf{x}_t^B$和$\mathbf{y}\tilde{t}$ 之前的系数矩阵分别记作 L 和 $K_{\tilde{t}}$，从而得到

$$\mathbf{x}_t \propto L\mathbf{x}_t^B + \sum_{\tilde{t}=t}^{t+\Delta t} K_{\tilde{t}}\mathbf{y}_{\tilde{t}} \tag{9}$$

需要说明的是，矩阵 $L = (I + \sum(\mathbf{HM}\dots\mathbf{M})^T(\mathbf{HM}\dots\mathbf{M}))^{-1}$ 在不同运行（或时间步）中始终保持相同的形式，因为$\mathbf{H}$ 中的元素仅取决于不同位置之间的相对（欧几里得）距离，而这些距离在不同运行（或时间步）中不会改变；$\mathbf{M}$ 中的元素则是非线性形式$\mathbf{x}_{t+\delta t} = m(\mathbf{x}_t)$ 在 $\mathbf{0}$ 处对任意 t求得的一阶导数，因此仅取决于两个时间步之间的相对差值，而不取决于任何显式时间 t，所以在不同运行（或时间步）中同样不会改变。

对于给定的一次运行，$\tilde{t}$ 的范围是 $t \sim (t + \Delta t)$；根据 $\tilde{t}$，$y_{\tilde{t}}$ 前面的每个系数矩阵形式各不相同。例如，如果 $\tilde{t} = t$，那么系数矩阵 $K_{\tilde{t}}$ 的形式为$\mathbf{HM}$；如果 $\tilde{t} = t + \delta t$，那么系数矩阵$K_{\tilde{t}}$的形式为$\mathbf{HMM}$，等等。

D 证明

D.1 引理 3 的证明

证明。 根据引理 1，$\mathbf{X}_t^A$是$\mathbf{Y}_t$的可测函数，因此在给定时刻 t，按空间索引排列的集合$\{\mathbf{X}_{s_i}^A(t),\ i = 1,\ 2,\ \dots\}$ 中的每个元素都是 $\mathbf{Y}_t$ 的可测函数，因此它们各自都是定义在 $\mathbf{Y}_t$ 的同一样本空间 Ω 上的按空间索引排列的随机对象。 □

D.2 定理 1 的证明

证明。 根据引理 3，$\{\mathbf{X}_{s_i}^A(t),\ i = 1,\ 2,\ \dots\}$ 是一个空间随机过程；在给定时刻 t，$\omega_t \equiv \mathbf{y}t$ 是一个固定的样本元素，因此根据定义，再分析输出$\{\mathbf{x}_{s_i}^A(t) = \mathbf{X}_{s_i}^A(\omega_t; t) : i = 1,\ 2,\ \dots\}$ 是该空间随机过程的一个实现或样本路径。 □

D.3 第 4.2.2 节的证明

在第一种情形中，我们假设网格位置按一维顺序排列，即对于两个位置 s_i 和 ^{s}j，$j > i$，于是我们将可测可逆变换函数T: Ω → Ω 定义为 $T^{\Delta s}(\omega_{s_1},\ \omega_{s_2},\ \dots\dots) = (\omega_{s_2},\ \omega_{s_3},\ \dots\dots)$，其中 Δs 是 s_i 与 ^{s}j之间的欧几里得距离单位，$(j > i)$，因此，我们有 $\mathbf{x}_{s_1}^A = \mathbf{X}_{s_1}^A(\omega_{s_1})$，

$$\begin{aligned}
\mathbf{x}_{s_2}^A &= \mathbf{X}_{s_2}^A(\omega_{s_2}) = \mathcal{G}(\omega_{s_2}) = \mathcal{G}(T^{\Delta s}(\omega_{s_1})) = \mathcal{G}T^{\Delta s}(\omega_{s_1}) = \mathbf{X}_{s_2}(\omega_{s_1}) \\
\mathbf{x}_{s_3}^A &= \mathbf{X}_{s_3}^A(\omega_{s_3}) = \mathcal{G}(\omega_{s_3}) = \mathcal{G}(T^{\Delta s}(\omega_{s_2})) = \mathcal{G}(T^{\Delta s}(T^{\Delta s}(\omega_{s_1}))) = \mathcal{G}T^{2\Delta t}(\omega_{s_1}) = \mathbf{X}_{s_3}(\omega_{s_1}) \\
&\ \ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots
\end{aligned}$$

因此，从上述方程的最右侧可得，我们在给定时刻得到一组再分析输出；从上述方程的最左侧可得，在固定样本元素 ω_{s_1}处，空间演化随机变量的一个实现。