ICS
CCS

# 团　　体　　标　　准

T/ACEF □□□—202X

# 河湖水环境数据同化技术指南

Technical guide for multisource data assimilation of river and lake water environment

（征求意见稿）

20□□-□□-□□发布　　20□□-□□-□□实施

中 华 环 保 联 合 会 发 布

# 目　次

# 前　言

本文件按照GB/T1.1-2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件为首次发布。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国水利水电科学研究院提出。

本文件由中华环保联合会归口。

本文件起草单位：中国水利水电科学研究院、

本文件主要起草人：

# 河湖水环境数据同化技术指南

## 1 范围

本文件规定了河湖数据同化算法选择、同化类型、同化流程和数据同化效果评估等技术内容。

本文件适用于河湖水环境监测数据处理，边界及初始数据驱动的水环境数值模型驱动、河湖水环境预警、预报，为提高数值模型的预测精度。

## 2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB 8566　　计算机软件开发规范

SL 219　　水环境监测规范

HJ 630　　环境监测质量管理技术导则

HJ 212　　污染物在线监控（监测）系统数据传输标准

HJ 355　　水污染源在线监测系统($COD_{Cr}$、$NH_3$-N 等)运行技术规范

## 3 术语和定义

下列术语和定义适用于本文件。

3.1

**监测及数据采集** monitoring and data collection

基于监测站网、物联网等，进行多源、全天候数据数据采集，实现数据实时、安全的传输和存储，进行高效的数据处理，提供多种场景的数据搜索，及延展的云处理等功能。

3.2

**数据融合** data fusion

在河湖水环境监测中，将空间、时间上频次及密度不一致的人工监测、自动传感监测、卫星遥感等多途径来源的水环境数据和信息加以联合、相关及组合的处理过程。

3.3

**数据同化** data assimilation

在考虑数据时空分布以及观测场、初始场或背景场误差的基础上，在数值模型的动态运行过程中融合新的观测数据的方法。它是在过程模型的动态框架内，通过数据同化算法不断融合时空上离散分布的不同来源和不同分辨率的直接或间接观测信息来自动调整模型轨迹，以改善动态模型状态的估计精度，提高模型预测能力。

3.4

**数值模型** numerical model

根据物质守恒、生化反应、物质转化等原理及过程，用数学语言和方法描述河湖物质在水体中发生的物理、化学、生物、化学和生态学诸方面的变化，包括在水体中发生的混合和输运、在时间和空间上的迁移转化、以及各因素相互作用关系的数学方程及求解。

3.5

**初始条件** Initial condition

初始条件是描述整个模型初始状态的数学表达，模型驱动后初始条件被新状态所替代，但初始条件需要接近模型计算的真实状态，确保模型的平稳驱动。

3.6

**边界条件** Boundary condition

边界条件是指模型在求解区域边界上的变量随时间和地点的变化规律，模型边界条件可以为给定数值边界、梯度边界、函数边界、自由边界等。

3.7

**模型参数** uncertainty input

数值模型运行计算需要输入边界条件、初始条件、模型计算参数等，进行方程求解，受观测条件描述过程公式等的限制，这几类数据的输入均可产生不确定性，导致模型计算结果的偏差。

3.8

**不确定性输入** uncertainty input

数值模型运行计算需要输入边界条件、初始条件、模型计算参数等，进行方程求解，受观测条件描述过程公式等的限制，这几类数据的输入均可产生不确定性，导致模型计算结果的偏差。

3.9

**误差分析** (Error Analysis )

误差分析是对模型预测结果可能存在的错误或不确定性的种类和数量的研究，分析模型结果与观测场的产生的偏移程度，以量化模型计算精度和数据同化效果。

## 4 数据同化算法

数据同化算法可划分为以下类型：

——最优插值法（Optimal interpolation，OI）

——粒子滤波算法（Particle Filter，PF）

——卡尔曼滤波算法（Kalman Filter，KF）

——集合卡尔曼滤波算法（Ensemble Kalman Filter，EnKF）

——变分算法（Variational Algorithm，VAR）

——层次贝叶斯方法（Hierarchical Bayesian Method，HBM）

——鲁棒滤波方法（HFilter，HF）

按数据同化算法与模型之间的关联机制，可将数据同化算法可分为顺序数据同化算法和连续数据同化算法两大类。

根据水环境数值模型求解的变量的时空特性，以及观测数据类型确定采用不同的同化算法。采用连续数据同化算法，根据观测数据时间确定同化的时间T，利用该同化时间的所有观测数据和模型状态变量值进行最优估计，通过迭代而不断调整模型初始场，最终将模型轨迹拟合到在同化时间上获取的所有观测场上，该类算法有三维变分、四维变分算法、集合卡尔曼滤波算法等。

采用顺序数据同化算法又称滤波算法，包括预测和更新两个过程。预测过程根据j时刻状态变量的值初始化模型，不断向前计算，直到有新的观测值输入，预测 j+1 时刻模型的状态变量值；更新过程则是对当前 j+1 时刻的观测值和模型状态预测值进行加权，得到当前时刻状态最优估计值。根据当前 j+1 时刻的状态值对模型重新初始化，重复上述预测和更新两个步骤，直到完成所有观测数据时刻的状态预测和更新，常见的算法有集合卡尔曼滤波和粒子滤波算法等。

## 5 数据同化类型

模型运算时，由于不确定性的输入，同化类型可分以下为三种类型

——初始场的同化：当模型计算值与实测值产生偏差，采用实测值对计算场进行同化，带入下随后的计算中，纠正模型计算的整体误差。对初始场进行同化时，应选择三维变分和四维变分算法、层次贝叶斯算法等相应同化算法。

——边界条件的同化：为克服边界输入的不确定性，确定需要同化的输入边界条件，可以是单一边界,或是多边界,采用同化技术对边界条件序列同化,使得计算结果逼近观测值，提高计算精度。对边界条件进行同化时，选择集合卡尔曼滤波和粒子滤波算法等相应算法。

——模型参数的同化：水环境模型通常计算参数众多，多参数带来参数的不确定性，对相关敏感性参数进行同化，该方法体现参数在不同时间段上的变化。对参数进行同化也需选择适当的同化算法，通常有集合卡尔曼滤波和粒子滤波算法等相应算法。

通过数值模型输入的不确定特征，选择某种同化类型并实施。

## 6 数据同化技术流程

### 6.1 数据收集整理

首先，根据所建立的水环境数学模型，收集整理需要驱动模型计算的数据过程或空间分布场，包括地形数据、气象数据、水文数据、水质指标的常规监测数据、自动监测站数据、卫星遥感解译等数据等，尽量减小模型输入的不确定性。其次，收集模型计算区域中状态变量的观测数据，用于数据同化。

应对于不同来源和类别的数据，按照空间和时间的分布密度、间隔时段等差异特点进行归类整理。

### 6.2 数据融合

尽可能采用自动化监测设备等自动数据采集的数据,从数据的一致性上做好数据的统一校核和整理分析，通过融合算法，如贝叶斯算法、最小二乘算法等，去除异常及故障数据，消除冗余，结合人工监测降低传感器系统偏差，保证数据准确性。

通过数据融合将多源数据进行科学、合理的综合处理，提高状态监测精度和数据诊断智能化程度，真实反映水环境状态及演变。

### 6.3 数值模型构建

根据取得资料和数据，及研究或预测需求，确定模型预测的状态变量，了解变量之间的相互影响与变化规律，选择可描述变量之间关系，反映现象的基本特征的过程模型，基

本框架由水动力模型耦合水温模型、水质模型等构成，设定研究时间段，和研究区域，状态变量的初始场，变量之间相互作用或影响的参数德国，以及输入驱动模型的边界过程等，开始模型的计算。

模型参数值确定可采用经验公式、室内实验或数学方法等，参数代入模型后能较好地变量之间的反应作用关系及过程，模型计算结果能较好重现研究区域水环境变化，基本符合观测数据。

## 6.4 数据同化

数值模型可预测多个变量的变化，包括水动力变量（水位、流量、流速等）、水温、水质指标的状态变量等。首先，确定模型输出的需要同化的预测变量，变量可以是一个也可是多个，同化变量需要在模型预测时间段内，有相应的观测值。在模型计算的时间轴上，判断是否有需要同化的模型求解变量的观测值，当该时段存在观测值时，调用同化算法进行同化计算，更新变量，或相应的模型参数、输入条件等，该时刻的同化结束后，采用更新后的数据继续计算，模型行进中，进行观测值判断，确定是否启动同化计算，直到模型计算结束。

通过同化技术可实现单一变量的同化，也可对多变量进行同时同化。在具体的流程见图6-1。

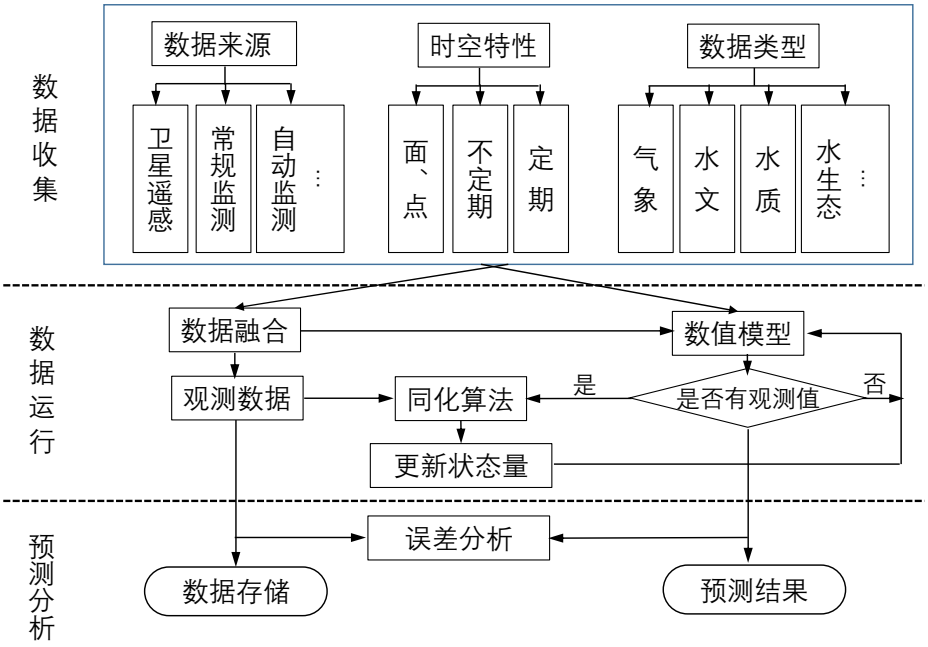


图 6-1. 数据同化流程

## 7. 数据同化效果评估

河湖水环境数据同化实施后，需对数据同化前后状态变量的误差进行评估，设定状态变量的观测数据为真实值，采用以下一种或几种不同的方法计算误差并分析，检验数据同化算法的有效性。若进行同化后，精度未能提升，应更换同化算法。

7.1 平均绝对误差(MAE)

用来衡量预测值与真实值之间的平均绝对误差，MAE 值越小表示模型预测越准确。

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,\ \in [0, +\infty)$$

7.2 均方误差（MSE）

平均的预测的值和真值差的平方，平均到每一个预测的值，MSE 值越小表示模型预测越准确。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,\ \in [0, +\infty)$$

7.3 均方根误差(Root Mean Square Error，RMSE)

亦称标准误差，是预测值与真实值偏差的平方与观测次数n比值的平方根，用来衡量观测值同真值之间的偏差。

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},\ \in [0, +\infty)$$

7.4 确定系数$R^2$

$R^2$ 对结果进行了归一化，更容易看出模型间的差距。$R^2$ 越接近1越好。

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \in [0, 1]$$

7.5 纳什效率系数（Nash-Sutcliffe efficiency coefficient，NSE）

一般用以验证水文、水环境模型模拟结果的的好坏。

$$\mathrm{NSE} = 1 - \frac{\Sigma_{t=1}^{T}\left(y_i^t - \widehat{y_i^t}\right)^2}{\Sigma_{t=1}^{T}\left(y_i^t - \bar{y}\right)^2}$$

对于一个完美的模型，估计的误差的方差等于0，则 NSE=1；相反，模型产生的估计误差方差等于观察到的时间序列的方差，结果NSE=0。实际上，NSE=0 表示该模型具有与时间序列平均值相同的预测能力，即误差平方和。当预测模型得到的估计误差方差显著大

于观测值方差时，NSE<0。NSE值越接近1，表明模型预测能力越好。因此NSE的取值范围为（-∞, 1]。但是如果将NSE用于模型回归中，则和R'完全等价，范围是 [0，1]。

以上各式中，$y_i$ 为观测值；$\hat{y}_i$为计算值；n为个数；t为有观测值的时间；$y^t$为 t 时刻的值。

# 附 录 A
（资料性）
数据同化算法

数据同化起源于20世纪50年代气象预报中对客观分析的需要，随着计算机技术不断发展。以下总结了几种数据同化技术算法及特点。

## A．1 最优插值法

最优插值（Optimal interpolation，OI）法源于20世纪90年代，主要有卫星高度计资料和常规温度的同化，是一种均方差最小的线性插值法，其原理是利用权重考虑背景场和观测误差的统计误差的统计特征。它能够处理不同精度的观测资料，在分析变量中考虑了彼此的线性关系，简单易行，计算量小。其缺点是背景协方差矩阵B是静态，综合比对时需要各种参数的经验与假设，所以该方法不是最优方法。

## A．2 粒子滤波算法

粒子滤波（Particle Filter，PF）是采用状态空间一组加权随机样本粒子逼近模拟变量的概率密度分布，随着粒子数目逐渐趋于无穷大时，粒子的概率密度函数逐渐逼近模拟变量的真实概率密度函数。它适用于非高斯非线性模型，是当前热门的数据同化算法。随着科技的发展，粒子滤波算法得到了不断地完善。通过不断选取合理的重要概率分布，增加变量粒子群的多样性，减少权重方差。

## A．3 卡尔曼滤波算法

卡尔曼滤波算法（Kalman Filter，KF）是1960年美国学者 Kalman 最早提出的，它是顺序数据同化算法的最早形式，同时也是数据同化算法的理论基础。随着时间的推移，卡尔曼滤波算法逐渐演化成顺序数据同化算法，此算法是从包含误差的数据资料里对所求变量利用最小二乘估计、最小方差估计等方法进行最佳估计，是以“预测—实测—预测”的顺序进行递推，使得状态估计值的误差降低至最小，从而得到状态量的最优估计。其缺点是状态估计值的结果依赖各参数间的误差，而参数误差的大小很难确定；它只能在状态线性变化和高斯假设分布条件下获得最优解，在实际中，状态量往往是非线性变化，这使得卡尔曼滤波的使用范围受到了限制。采用卡尔曼滤波算法实现数据同化分为2步：预测和更新，预测是依据当前 $j$-1 时刻模型的状态量预测j时刻的状态量：

$$x_j^f = f\left(x_{j-1}^a\right) \tag{1}$$

式中，$f$ 为非线性模型，它还量化了来自前一时间步长的估计的不确定性（即，系统误差或模型误差协方差，$P_j$）。在分析步骤期间，使用如下观测 $y_j$ 获得更新变量的向量：

$$x_j^a = x_j^f + K_j\left[y_j - Hx_j^f\right] \tag{2}$$

这里，$H$ 是将更新变量的向量映射到观测空间中的观测算子，$K_j$ 是定义为：

$$K_j = P_j H_j^T\left[H_j P_j^b H_j^T + R_j\right]^{-1} \tag{3}$$

更新是在已有观测数据的条件下对j时刻的状态预测值进行调整，从而得到 $j$ 时刻状态量的最优估计值。然后，利用 $j$ 的状态量估计值重新初始化模型，重复上述步骤，直至所有的观测数据都完成状态值预测与更新。

**A．4 集合卡尔曼滤波算法**

集合卡尔曼滤波（Ensemble Kalman Filter，EnKF）算法是 Evensen 首先提出来的，它是20世纪90年代中期集合预报与卡尔曼滤波的结合，EnKF用从模型运行的集合中获得的样本协方差矩阵来近似模型误差协方差矩阵。每个集合包括一组可能的模型轨迹，并且它们的分布用作信息源以找到卡尔曼滤波器方程（3），使用的更新状态变量 P 的协方差矩阵。引入集成卡尔曼滤波器来缓解与方程中模型误差协方差矩阵 P 的确定相关的概念和计算问题。

集合卡尔曼滤波不需要时间上的递推，通过统计方式计算状态的误差协方差矩阵，采用集合的思想解决了实际中背景误差协方差矩阵的估计和预报困难的难题，利用非线性系统的数据同化有效降低了数据同化计算量。

集合卡尔曼滤波算法是由预测和更新两个步骤组成。其优点是无需伴随模型，可应用于复杂的非线性系统；缺点是计算成本高，背景误差协方差的准确估计需要和模式纬度相当的集合样本量，集合样本较小时会引入明显的抽样误差。

**A．5 变分算法**

自20世纪80年代末以来，这些方法已被积极用于更新状态变量（Lawless 2013）。

变分算法起源于20世纪80年代，其基本思想是构建目标函数描述状态变量分析值和真值之间的差异，把数据同化转化为一个极值求解的问题[4]。变分算法的优点在于它引入了非线性观测算符和一些动力约束条件，打破了观测量和分析量之间存在线性关系的限制。

变分算法的典型代表是三维变分和四维变分，三维变分在构建目标函数中包含物理过程，并且是以模型预报值作为状态的背景场；四维变分的“四维”是指状态量空间三维分布和时间一维分布，它是在三维变分同化的基础上发展起来的。

三维变分算法（Three-Dimensional Variational Algorithm，3DVAR）是假设某一同化时刻t以及时间段T，用［*t*-T，*t*+T］时间段内所有观测数据调整模型轨迹，最终将模型轨迹拟合至该时间范围内所有观测值上。该方法的目的是找到更新变量（$x_0$）的值，该值最小化到背景的加权最小二乘距离（未更新的建模结果）更新变量 $x_b$ 加上到同化窗口中测量的加权最小二乘距。最小化的成本函数为：

$$\mathcal{J}\ (x_0) = \frac{1}{2}(x_0 - x_b)^T B^{-1}(x_0 - x_b) + (Hx_0 - y)^T\ R_i^{-1}(Hx_0 - y) \quad (3)$$

式中，*J(x)*是目标函数，x是状态量，xb是背景场，B是背景场误差协方差矩阵，Y是观测数据，H 是观测算子，R是观测场误差协方差矩阵，*J*(*x*) 取得最小的状态量为最有估计状态量。求 *J*(*x*) 最小值可以在转化为 *J*(*x*) 导数为0的状态值，对 *J*(*x*) 求一阶导数得到梯度方程。常用的优化算法有梯度下降法、牛顿法、松弛法、最优迭代步长法等。选择一种适当的优化算法，通过迭代得到 *J(x)* 的最优解。三维变分算法的缺点是设计一个合理的背景误差协方差矩阵 B 的模型比较困难，除此之外，三维变分算法的解在时间上是不连续的。

四维变分算法（Four- Dimensional Variational Algorithm，4DVAR）是由Talagrand在1987年提出，“四维”是指状态量空间三维分布和一维时间分布，是3DVAR的推广。它是利用时间连续的观测资料来优化初始时刻的模式状态场或模式参数，在一定程度上弥补了三维变分算法在状态量的时间变化以及初始化角度方面存在的不足之处。4DVAR目标函数的一般表达形式为：

$$J\left(x_0\right) = \frac{1}{2}\left(x_0 - x_b\right)^T B^{-1}\left(x_0 - x_b\right) + \frac{1}{2}\sum_{i=0}^{n}(y_i - H_i x_i)^T R_i^{-1}(y_i - H_i x_i) \quad (4)$$

公式（4）中*i*表示观测时刻；n表示同化窗口内的时间维总观测次数；$R_i$ 是第 $i$ 时刻的观测误差协方差矩阵；*Hi*是第*i*时刻的观测算子；$y_i$ 和 $x_i$ 分别代表第i时刻的观测和状态，$x_i$满足方程$x_i = \mathrm{M}_i$（$x_i$-1）；$M_i$ 是 $i$-1 时刻到i时刻的模式预报算子。

4DVAR 最重要的是建立伴随模型，常用的方法是差分的伴随，即由离散的正模型直接导出离散的伴随模型及目标函数的梯度表达式。它的优点是可以选择不同的控制变量，可以同化多时刻的资料，B 矩阵在同化窗口是隐式发展的，可以在目标函数中加上其他的弱约束。

**A．6 层次贝叶斯方法**

层次贝叶斯（Hierarchical Bayesian Method，HBM）方法是基于条件概率分布将复杂问题逐级分解成不同层次，各层次之间通过条件概率关联，把复杂的联合概率分布求解问题转化成一系列简单后验概率求解问题。

层次贝叶斯方法在数据同化过程中分为数据、过程和参数3个层次，数据由站点观测数据、动力模型输出数据以及遥感反演数据三部分组成，过程代等待同化参量的时空分布，参数包含有数值模型的所有参数。假设数据、过程和参数是3种随机变量，分别建立条件概率分布模型。将数据同化问题转化为已知数据条件下推理过程和参数的后验概率分布。

层次贝叶斯方法的优点是基于贝叶斯理论在已经给出观测数据和模型的条件下估计状态的后验概率；此方法充分考虑了数据同化中的不确定性以及不确定性对同化结果的影响，并且摆脱了线性和高斯假设的约束；采用分层次建模，逐级对各参数进行分析，分别建立相对应的条件概率模型，更加接近地球系统参量变化的客观规律；此方法是分层次以条件概率方式建立三层模型之间的联系，要求数据具有条件独立性，使得该方法更加符合理论和实际应用的要求。其缺点是实际应用中仅仅局限于小范围的数据同化实验中，对陆地数据同化研究较少；分层推理的形式使得前层多基于经验知识，没有固定理论，不合理的先验概率会影响后续结果的准确性；层次贝叶斯分层处理数据，目前研究主要集中在站点观测数据，对面-面数据同化研究较少。

**A．7 鲁棒滤波方法**

系统中含有不确定参数，精确的系统模型较难获得，为了解决这一问题，引入了鲁棒滤波方法，鲁棒滤波对可能的不确定性具有很好的容忍度。H滤波（HFilter，HF）是鲁棒滤波中的一种，这种滤波能够接受不完整的系统信息。

H 滤波将鲁棒控制设计中引入的性能指标 $H_\infty$ 范数应用于滤波，已解决系统中存在的各种不确定性问题，将噪声看作是能量有限的随机信号使得系统的干扰到估计误差的闭环传递函数的 $H_\infty$ 范数小于给定的正数。

## 参考文献

[1] 自然资源调查监测质量管理导则（试行）
[2] 国家地表水自动监测系统通讯协议技术导则（征求意见稿）