Received: 25 November 2022 | Revised: 25 May 2023 | Accepted: 25 July 2023

DOI: 10.1002/nla.2529

RESEARCH ARTICLE

WILEY

# Impact of correlated observation errors on the conditioning of variational data assimilation problems

**Olivier Goux**[1,2] | **Selime Gürol**[1] | **Anthony T. Weaver**[1] | **Youssef Diouane**[3] | **Oliver Guillet**[4]

[1]ALGO-COOP Team, CERFACS / CECI CNRS UMR 5318, Toulouse, France

[2]ISAE-SUPAERO, University of Toulouse, Toulouse, France

[3]Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal, Quebec, Canada

[4]CNRM UMR 3589, Météo-France and CNRS, Toulouse, France

**Correspondence**
Olivier Goux, CERFACS / CECI CNRS UMR 5318, 42 avenue Gaspard Coriolis, 31057 Toulouse Cedex 01, France.
Email: goux@cerfacs.fr

**Funding information**
Copernicus Climate Change Serive, Grant/Award Number: C3S_601_INRIA; French National Programme LEFE/INSU

## Summary

An important class of nonlinear weighted least-squares problems arises from the assimilation of observations in atmospheric and ocean models. In variational data assimilation, inverse error covariance matrices define the weighting matrices of the least-squares problem. For observation errors, a diagonal matrix (i.e., uncorrelated errors) is often assumed for simplicity even when observation errors are suspected to be correlated. While accounting for observation-error correlations should improve the quality of the solution, it also affects the convergence rate of the minimization algorithms used to iterate to the solution. If the minimization process is stopped before reaching full convergence, which is usually the case in operational applications, the solution may be degraded even if the observation-error correlations are correctly accounted for. In this article, we explore the influence of the observation-error correlation matrix (**R**) on the convergence rate of a preconditioned conjugate gradient (PCG) algorithm applied to a one-dimensional variational data assimilation (1D-Var) problem. We design the idealized 1D-Var system to include two key features used in more complex systems: we use the background error covariance matrix (**B**) as a preconditioner (B-PCG); and we use a diffusion operator to model spatial correlations in **B** and **R**. Analytical and numerical results with the 1D-Var system show a strong sensitivity of the convergence rate of B-PCG to the parameters of the diffusion-based correlation models. Depending on the parameter choices, correlated observation errors can either speed up or slow down the convergence. In practice, a compromise may be required in the parameter specifications of **B** and **R** between staying close to the best available estimates on the one hand and ensuring an adequate convergence rate of the minimization algorithm on the other.

**KEYWORDS**

condition number, conjugate gradient, convergence rate, covariance operators, diffusion operators, nonlinear weighted least-squares

## 1 | INTRODUCTION

An important class of nonlinear weighted least-squares problems arises from the assimilation of observations in atmospheric and ocean models, a procedure known as *data assimilation*. In data assimilation, observations of the state of a system are combined with an *a priori* estimate of the state, called the *background*, to produce an optimal estimate of the state of the system, called the *analysis*. In *variational data assimilation*, the optimal estimate is obtained iteratively by minimizing a nonlinear weighted least-squares cost function that is the sum of two terms: one measuring the model fit to the background state (the background term $\mathcal{J}_b$); the other measuring the model fit to the observations (the observation term $\mathcal{J}_o$), subject to constraints (generally nonlinear) that relate the model state to the observations. The weighting matrices for $\mathcal{J}_b$ and $\mathcal{J}_o$ are defined by an estimate of the inverse of the background and observation-error covariance matrices ($\mathbf{B}^{-1}$ and $\mathbf{R}^{-1}$), respectively. Variational data assimilation is widely used for operational state estimation in meteorology and oceanography as it is a practical method for solving nonlinear least-squares problems when the dimensions of the state and observation vectors are huge (typically $10^6$ to $10^9$).

In variational data assimilation, the cost function is minimized approximately using a truncated Gauss–Newton (GN) algorithm[1] or incremental variational data assimilation as it is known in the data assimilation community.[2] This reduces the nonlinear problem to a sequence of linear sub-problems (quadratic cost functions), each of which is solved iteratively using a preconditioned conjugate gradient (PCG) method.[3] Standard implementations of PCG for data assimilation employ $\mathbf{B}$ as a first-level preconditioner,[3,4] which we refer to as $\mathbf{B}$-PCG hereafter. Besides significantly improving the conditioning of the Hessian matrix,[5] $\mathbf{B}$-preconditioning allows the $\mathbf{B}$-PCG algorithm to be formulated in a way that avoids the need to specify $\mathbf{B}^{-1}$ explicitly. This is important as most $\mathbf{B}$ formulations used in practice are not associated with convenient representations of $\mathbf{B}^{-1}$.

There is still a requirement to specify $\mathbf{R}^{-1}$, however. To simplify its specification, practical implementations of $\mathbf{R}$ tend to have relatively simple structural forms. In the extreme yet common case, $\mathbf{R}$ is taken to be a diagonal matrix, which amounts to assuming that the observation errors are uncorrelated. This is a poor assumption for certain observations, especially from satellites.[6,7] If observation-error correlations are neglected when they are known to be important then the solution of the weighted least-squares problem will result in a degraded (sub-optimal) analysis and poor exploitation of the assimilated data. To mitigate the former while still using a diagonal $\mathbf{R}$, observation data sets are either 'thinned' into a subset of observations or aggregated into 'super-observations' that have reduced error correlations.[8] Furthermore, the observation-error variances are often multiplied by an 'inflation' factor in order to prevent the analysis from overfitting observations that may still have a substantial component of correlated error. However, when observation error is correlated over distances similar to or greater than those of the background error, inflation can actually degrade the analysis.[9] While these methods can alleviate to some extent the inaccuracies associated with a diagonal $\mathbf{R}$, they still lead to sub-optimal solutions since potentially valuable observations are excluded and any remaining error correlations from the pre-processed observations are ignored.[10]

Several studies have examined the impact from using non-diagonal representations of $\mathbf{R}$ to account for spatially correlated errors.[11-14] A general conclusion that arises in most of these works is that accounting for spatial correlations in $\mathbf{R}$ leads to a more accurate solution, especially for the smaller spatial scales, even with a rather crude correlation model. However, even crude correlation models can lead to impractical representations of $\mathbf{R}^{-1}$. Various correlation models with accessible inverse representations have been proposed in the literature.[15-19] One of the challenges with specifying $\mathbf{R}$ and $\mathbf{R}^{-1}$ is that observation locations tend to be arranged in an arbitrary and unpredictable way due to the measurement method or quality control procedures that result in observations being removed. This means that correlation models developed for structured grids, like those typically associated with $\mathbf{B}$, are not necessarily applicable for $\mathbf{R}$.

In this study, we use diffusion operators to model spatial correlations in both $\mathbf{B}$ and $\mathbf{R}$. Diffusion operators can be used to model correlation functions from the Matérn class[20] and provide convenient and inexpensive representations of the associated inverse correlation operators.[21,22] They are popular for representing spatially correlated background error in complex boundary domains such as those encountered in ocean data assimilation.[4,23-26] Furthermore, Reference 18 describes how to adapt these operators to unstructured meshes and hence to make them suitable for $\mathbf{R}$ and $\mathbf{R}^{-1}$.

The rate of convergence of the conjugate gradient (CG) method is mainly determined by the characteristics of the eigenvalue spectrum of the Hessian matrix.[3,27] As the eigenvalues of the Hessian matrix are strongly dependent on $\mathbf{B}$ and $\mathbf{R}$, we can expect a non-diagonal $\mathbf{R}$ to have a significant impact on the rate of convergence of $\mathbf{B}$-PCG. In operational data assimilation, analyses must be delivered subject to strict computational constraints, which means that the stopping criterion for $\mathbf{B}$-PCG is usually set by a maximum allowed number of iterations rather than a measure of the convergence of the solution. Therefore, it is essential to ensure that the rate of convergence of $\mathbf{B}$-PCG from the use

of a non-diagonal **R** is not deteriorated to an extent that it outweighs the benefits brought from specifying a more accurate **R**.

In previous work, Reference 28 analysed the convergence rate for the special case of a diagonal **R** and discussed the influence on the condition number of observation and background accuracy, observation density, and the background-error correlation length-scale. Reference 29 studied the effects of a non-diagonal **R** on the convergence rate of the unpreconditioned CG method and derived general theoretical bounds for the condition number. These results were extended by Reference 30 to the **B**-preconditioned case (**B**-PCG). In both studies, theoretical and experimental results were obtained for the special case where the background- and observation-error correlation matrices are defined as circulant matrices. Their numerical experiments were performed using a particular circulant matrix built from a second order auto-regressive (SOAR) correlation function defined on the one-dimensional (1D) circular domain $\mathbb{S}$.

In this article, we are interested in understanding the sensitivity of the condition number with respect to the basic parameters of the diffusion operators that are used to model background- and observation-error correlations. Correlation functions derived from diffusion operators are controlled by a smoothness parameter $M$ (the number of diffusion iterations) as well as a length-scale parameter $L$ (the square-root of the diffusion coefficient), and thus are more flexible than the SOAR function, which is controlled by a single length-scale parameter. In fact, the SOAR function corresponds to a particular member ($M = 2$) of the family of correlation kernels represented by the 1D diffusion operator. We illustrate how the relative choice of $M$ for **B** and **R** can have a profound effect on the conditioning of the minimization problem.

The organization of the article is as follows. In Section 2, we introduce the weighted least-squares problem underlying variational data assimilation and we outline the solution algorithm based on truncated GN combined with CG. We provide the background theory on CG (and **B**-PCG) that is needed in this article for establishing the theoretical results and for interpreting the results from the numerical experiments with a 1D variational data assimilation (1D-Var) system. We conclude this section with a description of **B** and **R**, exposing the fundamental covariance parameters that control the shape characteristics of the correlation functions as well as the conditioning of the CG minimization. The formulation of **B** and **R** in terms of diffusion operators depends on theoretical results that are summarized in Appendix A. In Section 3, we study the eigenvalue spectrum of the Hessian matrix under the assumption that the observations are sampled from a regular grid and evenly distributed. While restrictive, this assumption allows us to derive analytical bounds on the condition number and to gain crucial insights on the convergence of **B**-PCG in more general frameworks. First, we present the general bounds that were derived by References 29 and 30. Then, we derive specific bounds that take into account the structural properties of the diffusion operators used to model **B** and **R**. We relegate the technical details of the proofs of a key theorem and associated corollaries to Appendices B,C, and D. In Section 4, we present the results from numerical experiments with the 1D-Var system to examine the sensitivity of the convergence of **B**-PCG to the parameters of **B** and **R**. These results show a strong sensitivity of the condition number, and hence convergence rate, to the correlation parameters. We argue that the parameter values should be chosen as a compromise between specifying the most accurate correlation model on the one hand and achieving a satisfactory convergence rate for the CG minimization on the other. We provide a summary and conclusions in Section 5.

## 2 | THE WEIGHTED LEAST-SQUARES PROBLEM

### 2.1 | Problem formulation: Variational data assimilation

Variational data assimilation provides an estimate of the physical state of a system by combining *a priori* information (the background state) and observations, together with information about their uncertainties. Here, we will use mathematical notation that is standard in meteorological and ocean data assimilation.[31] Assuming unbiased Gaussian error statistics for the background state and observations, the estimation problem can be formulated as a nonlinear weighted least-squares problem defined by the cost function

$$\min_{\mathbf{x}} \mathcal{J}(\mathbf{x}) = \frac{1}{2} \| \mathbf{x} - \mathbf{x}_\mathrm{b} \|^2_{\mathbf{B}^{-1}} + \frac{1}{2} \| \mathcal{H}(\mathbf{x}) - \mathbf{y}_\mathrm{o} \|^2_{\mathbf{R}^{-1}}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the state vector to be optimized and $\mathbf{x}_\mathrm{b} \in \mathbb{R}^n$ is the background estimate of the state vector. The vector of observations is $\mathbf{y}_\mathrm{o} \in \mathbb{R}^m$, and $\mathcal{H}(\cdot)$ is the observation operator, which maps an estimate of the state of the system to its equivalent in observation space. In general, $\mathcal{H}(\cdot)$ is nonlinear and non-bijective. In four-dimensional variational

assimilation (4D-Var), $\mathcal{H}(\cdot)$ would contain the forecast model operator, $\mathbf{x}$ would be the initial state vector, and $\mathbf{y}_o$ would be a vector that concatenates observations distributed over a given time window. The unbiased Gaussian distributions of the background and observation errors are characterized statistically by the covariance matrices $\mathbf{B} \in \mathbb{R}^{n\times n}$ and $\mathbf{R} \in \mathbb{R}^{m\times m}$, respectively. By definition, $\mathbf{B}$ and $\mathbf{R}$ are symmetric positive-definite (SPD) matrices. The inverse covariance matrices $\mathbf{B}^{-1}$ and $\mathbf{R}^{-1}$ define inner products in the background and observation spaces, and are used as weighting matrices in the cost function (1) where $\|\mathbf{x}\|_{\mathbf{P}}^2 = \mathbf{x}^{\mathsf{T}}\mathbf{P}\mathbf{x}$ denotes the squared $\mathbf{P}$-norm of a vector. The analysis is the global minimizing solution: $\mathbf{x}_a = \arg\min \mathcal{J}(\mathbf{x})$.

Truncated Gauss–Newton (GN),[1] which is known as incremental variational assimilation in the meteorological and ocean data assimilation communities,[2] is a common method for finding an approximate minimum of the nonlinear cost function (1). Truncated GN approaches the solution iteratively by solving, on each GN iteration $k$, the linearized sub-problem

$$\min_{\delta\mathbf{x}} J(\delta\mathbf{x}) = \frac{1}{2}\|\mathbf{x}_k - \mathbf{x}_b + \delta\mathbf{x}\|_{\mathbf{B}^{-1}}^2 + \frac{1}{2}\|\mathbf{H}_k\delta\mathbf{x} - \mathbf{d}_k\|_{\mathbf{R}^{-1}}^2, \tag{2}$$

which is a quadratic approximation of the non-quadratic cost function (1) in a neighborhood of the current iterate $\mathbf{x}_k$. In Equation (2), $\mathbf{H}_k \in \mathbb{R}^{m\times n}$ is the observation operator linearized about $\mathbf{x}_k$, and $\mathbf{d}_k = \mathbf{y}_o - \mathcal{H}(\mathbf{x}_k) \in \mathbb{R}^m$ is the misfit between the observation vector and the current iterate mapped to observation space. If $\delta\mathbf{x}_{k+1}$ denotes the solution of (2) then the estimate of the state is updated according to

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \delta\mathbf{x}_{k+1},$$

where $k = 0, \ldots, K-1$ and $\mathbf{x}_0 = \mathbf{x}_b$ (in general). In data assimilation applications with atmospheric or ocean models, the maximum number of GN iterations is typically very small ($K < 10$) for computational reasons.

The quadratic sub-problem (2) can be rewritten in standard quadratic form

$$\min_{\delta\mathbf{x}} J(\delta\mathbf{x}) = \frac{1}{2}\delta\mathbf{x}^{\mathsf{T}}\mathbf{A}_k\delta\mathbf{x} - \mathbf{b}_k^{\mathsf{T}}\delta\mathbf{x} + c_k, \tag{3}$$

where

$$\mathbf{A}_k = \mathbf{B}^{-1} + \mathbf{H}_k^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{H}_k,$$

is the SPD approximation of the Hessian matrix of the nonlinear cost function,

$$\mathbf{b}_k = \mathbf{B}^{-1}(\mathbf{x}_b - \mathbf{x}_k) + \mathbf{H}_k^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{d}_k,$$

is the negative gradient of the nonlinear cost function with respect to the current iterate $\mathbf{x}_k$, and $c_k = J(\mathbf{0})$ is a scalar. Satisfying the optimality condition of the quadratic sub-problem (3) requires solving the linear system

$$\mathbf{A}_k\delta\mathbf{x} = \mathbf{b}_k.$$

For our target applications, the dimension ($n$) of the state vector is large and the matrices are generally only available as operators (i.e., via matrix-vector products, not explicit matrices). For this reason, it is very common to solve the quadratic sub-problem iteratively using CG methods.

## 2.2 | Solving the quadratic sub-problem with the conjugate gradient method

Conjugate gradient (CG) is a Krylov subspace method (see Golub and Van Loan [Reference 32, Sec. 6.7] and Saad [Reference 33, Sec. 10.2]) for solving linear systems where the system matrix is SPD. CG seeks an approximate solution

$$\delta\mathbf{x}_\ell \in \delta\mathbf{x}_0 + \mathcal{K}^\ell(\mathbf{A}_k, \mathbf{b}_k),$$

where $\delta\mathbf{x}_0$ is the initial approximation and

$$\mathcal{K}^{\ell}(\mathbf{A}_k, \mathbf{b}_k) = \text{span}\{\mathbf{b}_k, \mathbf{A}_k\mathbf{b}_k, \ \dots \ , \mathbf{A}_k^{\ell-1}\mathbf{b}_k\},$$

is the Krylov subspace generated by $\mathbf{A}_k$ and $\mathbf{b}_k$. When $\mathbf{x}_0 = \mathbf{x}_\text{b}$, the initial iterate $\delta\mathbf{x}_0 = \mathbf{0}$. Hereafter, we will drop the truncated GN iteration index $k$ for clarity of notation. In order to find a unique solution, CG imposes the orthogonality condition

$$\mathbf{r}_\ell \perp \mathcal{K}^{\ell}(\mathbf{A}, \mathbf{b}),$$

where $\mathbf{r}_\ell = \mathbf{b} - \mathbf{A}\delta\mathbf{x}_\ell$ is the residual at the $\ell$-th iteration of CG. As a result, CG minimizes the quadratic cost function given by (3) over the subspace $\delta\mathbf{x}_0 + \mathcal{K}^{\ell}(\mathbf{A}, \mathbf{b})$ [Reference 34, theorem 5.2], so that the $\ell$-th iterate $\delta\mathbf{x}_\ell$ minimizes the error $\mathbf{e}_\ell = \delta\mathbf{x}^* - \delta\mathbf{x}_\ell$ in the $\mathbf{A}$-norm over the same Krylov subspace, $\delta\mathbf{x}^*$ being the exact solution [Reference 35, lemma 2.1.1]. The convergence properties of CG can then be analysed in terms of the error in the $\mathbf{A}$-norm [Reference 33, pp. 204–205]:

$$\frac{\|\mathbf{e}_\ell\|_\mathbf{A}}{\|\mathbf{e}_0\|_\mathbf{A}} \leq 2\left(\frac{\sqrt{\kappa(\mathbf{A})} - 1}{\sqrt{\kappa(\mathbf{A})} + 1}\right)^{\ell}, \quad \ell \in \mathbb{N}, \tag{4}$$

where $\kappa(\mathbf{A})$ is the condition number of $\mathbf{A}$, which is defined in the 2-norm as

$$\kappa(\mathbf{A}) = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})},$$

with $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ being the largest and smallest eigenvalues of $\mathbf{A}$, respectively. Equation (4) shows that convergence will tend to be fast when the condition number is close to 1. The condition number can thus be used as an indicator of the convergence rate of CG. Note that the initial error may also have an influence on the convergence behavior. For simplicity, however, we focus only on the effect of the condition number on the convergence rate.

Since Equation (4) depends on the condition number, it does not account for the distribution of the eigenvalues between the smallest and largest values. As a consequence, it can lead to a pessimistic error bound, especially when $\kappa$ is very large. More advanced error bounds exist, which account for a more complex representation of the spectrum (e.g., see Chap. 13 in the book of Reference 36). However, for the problem considered in this article, those error bounds provide little improvement over the error bound given by Equation (4). They have been shown to be more accurate only in the limit of a very large number of iterations, while here we are interested mainly in the solution accuracy in the early iterations of CG. As they are more complex and less general than Equation (4), we did not apply them in this article.

In order to accelerate the convergence rate of CG, it is common to use a preconditioner. For data assimilation problems that solve quadratic problem (3), it is customary to use $\mathbf{B}$ as a preconditioner, as it usually yields a significantly smaller condition number compared to that of the unpreconditioned problem, and a more clustered spectrum of eigenvalues.[3,5,37] Therefore, we will focus on solving the $\mathbf{B}$-preconditioned linear system. Since $\mathbf{B}$ is SPD, it can be factored as

$$\mathbf{B} = \mathbf{U}\mathbf{U}^\mathsf{T},$$

where $\mathbf{U} \in \mathbb{R}^{n\times n}$. We can then introduce $\mathbf{B}$-preconditioning symmetrically using a split-preconditioner,

$$\mathbf{U}^\mathsf{T}\mathbf{A}\ \mathbf{U}\ \delta\mathbf{v} = \mathbf{U}^\mathsf{T}\mathbf{b}, \tag{5}$$

where $\delta\mathbf{x} = \mathbf{U}\delta\mathbf{v}$. An unpreconditioned CG can be applied to Equation (5) by taking $\mathbf{U}^\mathsf{T}\mathbf{A}\mathbf{U}$ as the (SPD) system matrix and $\mathbf{U}^\mathsf{T}\mathbf{b}$ as the right-hand side.

In this article, we will evaluate the condition number of the preconditioned Hessian matrix,

$$\mathbf{S} = \mathbf{U}^\mathsf{T}\mathbf{A}\ \mathbf{U} = \mathbf{I}_n + \mathbf{U}^\mathsf{T}\mathbf{H}^\mathsf{T}\mathbf{R}^{-1}\ \mathbf{H}\ \mathbf{U}, \tag{6}$$

and determine its sensitivity to parameters in the covariance matrices $\mathbf{B}$ and $\mathbf{R}$ when their spatial correlations are modeled by diffusion operators.

## 2.3 | Weighting matrices formulated as the inverse of diffusion operators

The operators **B** and **R** describe the covariance structures of the background and observation errors. In the idealized 1D-Var system used in this study, the covariance matrices are small enough to be constructed explicitly using a functional expression to determine the matrix elements (e.g., as done in Reference 30). However, when one considers a realistic system, the size of **B** and **R** become too large to perform direct matrix-vector products. Hence, we prefer to adopt an approach that scales with the size of the problem and avoids the explicit construction of covariance matrices.

Covariance matrices can be factored as $\mathbf{\Sigma C \Sigma}$, where $\mathbf{\Sigma}$ is a diagonal matrix of standard deviations and **C** is an SPD correlation matrix. The computational difficulties are inherent in the specification and application of **C**. Reference 23 showed that multiplying an arbitrary vector by a Gaussian correlation matrix can approximately be achieved by numerically 'time'-stepping a diffusion equation with that arbitrary vector taken as the 'initial' condition[1]. This procedure defines an operator that models the product of a correlation matrix with the 'initial' condition without defining each element of the matrix. Since each 'time'-step involves the manipulation of sparse matrices, this strategy is naturally appropriate for large problems. The product of the diffusion coefficient $\mu$ and the total action 'time' $T = M\Delta t$ of the diffusion process, where $M$ is the total number of diffusion steps and $\Delta t$ is the 'time' step, controls the length-scale $D_g$ of the Gaussian function that would be used to construct the correlation matrix, where $D_g^2 = 2\mu T$. Reference 24 describe the technique in detail and generalize it to account for anisotropic correlations. References 21 and 22 describe an extension of the technique that involves solving the diffusion equation using an implicit 'time'-stepping scheme instead of the explicit scheme of the original approach.

With the implicit scheme, the total number of diffusion steps $M$ becomes a free parameter together with the diffusion coefficient multiplied by the 'time' step ($\mu\Delta t$ [With the explicit scheme, their product is the single free parameter controlling the length-scale $D_g$ of the Gaussian function]). This extra degree of freedom allows the diffusion operator to represent matrix-vector products with correlation matrices from the Matérn family,[20] where $M$ is linked to the standard smoothness parameter $\nu$ of the underlying Matérn correlation functions in $\mathbb{R}^d$ via the relation $\nu = M - d/2$. In $\mathbb{R}$, these functions are characterized by a polynomial times the exponential function and are also known as $M$th-order auto-regressive (AR) functions. The parameter $\mu\Delta t$ is precisely the square of the standard length-scale parameter $L$ of the Matérn functions. Without loss of generality, $\Delta t$ can be set to 1, so that $\mu = L^2$. The connection between the Matérn correlation functions and the differential operator describing the inverse of an implicit diffusion process has its roots in the seminal work of Reference 38. We outline the connection in Appendix A and provide the key theoretical relations for the 1D-Var problem under consideration in this study.

While $L$ is the length-scale parameter appearing explicitly in the definition of the diffusion operator, it is common in data assimilation to use an alternative length-scale parameter, the Daley length-scale $D$, to control the spatial smoothing properties of the diffusion kernel. The Daley length-scale can be understood as the half-width of the parabola osculating the correlation function at its origin.[39,40] It is defined for at least twice differentiable correlations functions, which in our case corresponds to AR functions with $M > 1$. As discussed in Appendix A.1, on $\mathbb{R}$, $D$, and $L$ are related through the equation

$$D = L\sqrt{2M - 3}, \tag{7}$$

where the square-root term generalizes to $\sqrt{2M - d - 2}$ in $\mathbb{R}^d$.[22] An advantage of the parameter pair $(D, M)$ over $(L, M)$ is that it allows better control of the spatial range and spectral properties of the AR functions (see Figure A1 in Appendix A.1). An important property is that AR functions converge to a Gaussian function as $M$ tends to infinity with $L$ simultaneously reduced to zero to keep $D$ constant. A closely-related length-scale parameter,

$$\rho = wL\sqrt{2M - 1}, \tag{8}$$

where $w$ is a constant, is used in geostatistical and machine-learning applications.[41-432] AR functions defined in terms of $\rho$ also have the property that they converge to a Gaussian function as $M$ tends to infinity with $\rho$ fixed. In $\mathbb{R}^d$, the square-root term in Equation (8) generalizes to $\sqrt{2M - d}$. An advantage of $\rho$ over $D$ is that it is valid for $M = 1$ as well as $M > 1$, while an advantage of $D$ over $\rho$ (and $L$) is that it is easier to estimate in practical applications when $D$ is spatially dependent [Reference 26, Sec. 2.4]. In Section 3, the analytical results are first derived in terms of $L$ and then interpreted in terms of both $D$ and $\rho$, where we will refer to $\rho$ with $w = 1$ as the *Stein* length-scale for convenience. In Section 4, the numerical experiments are discussed mainly in terms of $D$.

On a circular domain of radius $a$ (see Appendix A.2), we can define the discrete, symmetric diffusion-modeled correlation operator $\mathbf{C}$ as a sequence of linear operators represented by their respective matrices [Reference 25, Sec. 3.1]:

$$\mathbf{C} = \mathbf{\Gamma}\,\mathbf{L}\,\mathbf{W}^{-1}\,\mathbf{\Gamma}, \tag{9}$$

where $\mathbf{L} = \mathbf{T}^{-M}$ is a self-adjoint diffusion operator, $\mathbf{T}$ being a discrete representation of the shifted Laplacian operator $\mathcal{T}$. On the circular domain with constant $L$, we have from Equation (A5) that $\mathcal{T} \equiv I - L^2\partial^2/a^2\partial\phi^2$ where $-\pi \le \phi \le \pi$. The matrix $\mathbf{W}$ contains geometry- and grid-dependent weights. It defines the weighting matrix of the discrete form of the $L^2(\mathbb{S})$-inner product with respect to which $\mathbf{L}$ is self-adjoint; that is, $\mathbf{L} = \mathbf{W}^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{W}$. Sections 3.2 and 3.3 of Reference 18 provide a comprehensive discussion of this point within the context of a finite element method discretization of the diffusion equation. The diagonal matrix $\mathbf{\Gamma}$ contains normalization factors so that the diagonal elements of $\mathbf{C}$ are approximately equal to one. On the circular domain with constant $L \ll a$, we can set $\mathbf{\Gamma} = \gamma\mathbf{I}$ where $\gamma$ is well approximated by the constant product $\sqrt{\nu L}$ where $\nu$ is a monotonically increasing function of $M$ given by Equation (A3). For example, the error in $\gamma^2$ is smaller than 0.001% when $L/a = 0.3$. Reference 26 provide an overview of methods for estimating $\mathbf{\Gamma}$ on other domains and when the correlation parameters are not constant.

Taking $M$ to be an even number allows us to split $\mathbf{T}^{-M} = \mathbf{T}^{-M/2}\,\mathbf{T}^{-M/2}$ and hence to derive a simple 'square-root' factorization of Equation (9). The 'square-root' operator is convenient for generating random correlated samples and has been used for this purpose for the numerical experiments in Section 4. Another convenient property that comes specifically from the implicit formulation is that it provides immediate access to an inexpensive formulation of the inverse of the operator. This can be noticed from the inverse of Equation (9),

$$\mathbf{C}^{-1} = \mathbf{\Gamma}^{-1}\,\mathbf{W}\,\mathbf{L}^{-1}\,\mathbf{\Gamma}^{-1},$$

where $\mathbf{L}^{-1} = \mathbf{T}^{M}$ simply involves $M$ applications of the shifted Laplacian operator.

Here, we consider a finite-difference discretization of the diffusion equation under the assumption that the grid resolution is uniform so that $\mathbf{W} = h\mathbf{I}$, where $h$ is the grid size. Furthermore, we assume that $\mathbf{\Sigma} = \sigma\mathbf{I}$ where $\sigma$ is a constant standard deviation. Given these assumptions together with the assumption that $L$ is constant, we can simplify the expressions for the diffusion-modeled covariance operators for $\mathbf{B}$ and $\mathbf{R}$ as

$$\mathbf{B} = \frac{\sigma_{\mathrm{b}}^2\nu_{\mathrm{b}}L_{\mathrm{b}}}{h_{\mathrm{b}}}\mathbf{T}_{\mathrm{b}}^{-M_{\mathrm{b}}} = \frac{\sigma_{\mathrm{b}}^2\nu_{\mathrm{b}}L_{\mathrm{b}}}{h_{\mathrm{b}}}\left(\mathbf{I}_n - L_{\mathrm{b}}^2\mathbf{\Delta}_{h_{\mathrm{b}}}\right)^{-M_{\mathrm{b}}}, \tag{10}$$

$$\mathbf{R} = \frac{\sigma_{\mathrm{o}}^2\nu_{\mathrm{o}}L_{\mathrm{o}}}{h_{\mathrm{o}}}\mathbf{T}_{\mathrm{o}}^{-M_{\mathrm{o}}} = \frac{\sigma_{\mathrm{o}}^2\nu_{\mathrm{o}}L_{\mathrm{o}}}{h_{\mathrm{o}}}\left(\mathbf{I}_m - L_{\mathrm{o}}^2\mathbf{\Delta}_{h_{\mathrm{o}}}\right)^{-M_{\mathrm{o}}}, \tag{11}$$

where the subscripts 'b' and 'o' refer to quantities relative to the background and observations, respectively. The symbol $\mathbf{\Delta}_h$ denotes the finite-difference representation of the Laplacian operator, which depends on the grid resolution for the background and observations. The matrices $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ and $\mathbf{I}_m \in \mathbb{R}^{m\times m}$ are identity matrices. We are also interested in the expression for $\mathbf{R}^{-1}$, which follows immediately from Equation (11):

$$\mathbf{R}^{-1} = \frac{h_{\mathrm{o}}}{\sigma_{\mathrm{o}}^2\nu_{\mathrm{o}}L_{\mathrm{o}}}\mathbf{T}_{\mathrm{o}}^{M_{\mathrm{o}}} = \frac{h_{\mathrm{o}}}{\sigma_{\mathrm{o}}^2\nu_{\mathrm{o}}L_{\mathrm{o}}}\left(\mathbf{I}_m - L_{\mathrm{o}}^2\mathbf{\Delta}_{h_{\mathrm{o}}}\right)^{M_{\mathrm{o}}}.$$

By taking $\mathbf{W}_{\mathrm{o}} = h_{\mathrm{o}}\mathbf{I}_m$, we are assuming that the observations are regularly distributed with a separation distance of $h_{\mathrm{o}}$. This is done for mathematical convenience. Relative to the domain size $2\pi a$, $h_{\mathrm{o}}$ is an explicit parameter that reflects observation density and can be compared to $h_{\mathrm{b}}/2\pi a$, the density of background points. With the simplifying assumptions above, we are able to establish explicit theoretical bounds on the condition number of the preconditioned Hessian matrix as detailed in Section 3. Note that these assumptions, while necessary for establishing theoretical results, are not met in realistic atmospheric or ocean applications. In particular, most observation types have unstructured spatial distributions. Nevertheless, we can expect the results derived in Section 3, which focus on the sensitivity of the condition number of the Hessian matrix to correlation model parameters, to be relevant even with more complicated observation networks.

We remark now on the actual values of the parameter pairs $(M_{\mathrm{b}}, M_{\mathrm{o}})$, $(L_{\mathrm{b}}, L_{\mathrm{o}})$, and $(\sigma_{\mathrm{b}}^2, \sigma_{\mathrm{o}}^2)$ that will be considered in this study. First, values of $(M_{\mathrm{b}}, M_{\mathrm{o}}) \ge 10$ lead to AR functions that are practically Gaussian, so we will not consider values

beyond 10. Values of $(L_b, L_o)$ should be large enough compared to the grid size $(h_b, h_o)$ (at least $L_b/h_b \geq 1$ and $L_o/h_o \geq 1$) in order to avoid large discretization errors in the finite-difference representation of the diffusion operator. Ideally, the parameters should be chosen to provide the optimal fit to our available estimate of the error covariances (with $(L_b, L_o)$ and $(\sigma_b^2, \sigma_o^2)$ made spatially dependent in general). Background-error correlations are often specified as quasi-Gaussian functions (large values of $(M_b, M_o)$). The reason for this choice can be mainly computational; that is, efficient models, like diffusion, exist for applying quasi-Gaussian functions.[24,44,45] Another reason is that quasi-Gaussian functions are sufficiently regular that they can be differentiated, which is important for defining cross-variable (multivariate) covariances in atmospheric and ocean data assimilation.[39,46,47] In comparison, estimates of the spatial correlations of observation error often display a sharp decrease at short range and slow decay at longer range,[7,16,48] which with an AR function is best modeled with a small value of $(M_b, M_o)$. In view of these remarks, the case where $M_o < M_b$ seems to be of particular interest. Nevertheless, both this case and the case where $M_o \geq M_b$ will be considered as different data-sets may give rise to different error characteristics.

## 3 | CONDITIONING OF THE PRECONDITIONED LINEAR SYSTEM

In this section, we are interested in analysing the convergence of CG applied to the linear system (5) where the system matrix $\mathbf{S}$ depends on the diffusion-modeled covariance matrices described in the previous section. In particular, we are interested in analysing the sensitivity of the convergence in terms of the parameters of these covariance matrices. For this purpose, we will focus on the condition number of $\mathbf{S}$, denoted $\kappa(\mathbf{S})$.

We start by recalling some results from References 28 and 30 on the upper bound of $\kappa(\mathbf{S})$ for general covariance matrices.

**Theorem 1** (Theorem 3 of Reference 30). *Let $\mathbf{B} \in \mathbb{R}^{n\times n}$ and $\mathbf{R} \in \mathbb{R}^{m\times m}$ be symmetric, positive-definite matrices. Let $\mathbf{U} = \mathbf{U}^T \in \mathbb{R}^{n\times n}$ be the (unique) symmetric square root of $\mathbf{B} = \mathbf{U}\mathbf{U}^T = \mathbf{U}^2$ and let $\mathbf{V} = \mathbf{V}^T \in \mathbb{R}^{m\times m}$ be the (unique) symmetric square root of $\mathbf{R} = \mathbf{V}\mathbf{V}^T = \mathbf{V}^2$. If $\mathbf{H} \in \mathbb{R}^{m\times n}$, with $m < n$, and $\mathbf{S} = \mathbf{I}_n + \mathbf{U}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{U}$, then*

$$\kappa(\mathbf{S}) \leq 1 + \|\mathbf{V}^{-1}\mathbf{H}\mathbf{B}\mathbf{H}^T\mathbf{V}^{-1}\|_\infty. \tag{12}$$

The upper bound given in Equation (12) is quite general and it is not straightforward to understand the impact of each component of the matrix $\mathbf{S}$ on the condition number. Moreover, caution is needed when applying Equation (12) in physical applications as the eigen-decomposition of a covariance matrix will not be independent of the physical units of the variables [Sec. 4.3.4; Reference 49]. For this reason, it is generally more meaningful to consider an eigen-decomposition on the associated (non-dimensional) correlation matrix. In our idealized study, there is a single "physical" variable, with arbitrary units, and direct observations of that variable. Furthermore, both the background- and observation-error variances are taken to be constant, so can be factored out of $\mathbf{B}$ and $\mathbf{R}$. In this case, the eigenvalues of $\mathbf{B}$ and $\mathbf{R}$ will be identical to those of their respective correlation matrices up to a multiplicative factor given by their respective variances. We can thus continue to consider the eigen-decompositions of $\mathbf{B}$ and $\mathbf{R}$ without ambiguity.

Other bounds have been proposed by References 28 and 30 to understand the impact of each covariance matrix.

**Theorem 2** (Corollary 1 of Reference 30). *Let $\mathbf{B} \in \mathbb{R}^{n\times n}$ and $\mathbf{R} \in \mathbb{R}^{m\times m}$ be symmetric, positive-definite matrices. Let $\mathbf{U} = \mathbf{U}^T \in \mathbb{R}^{n\times n}$ be the (unique) symmetric square root of $\mathbf{B} = \mathbf{U}\mathbf{U}^T = \mathbf{U}^2$. If $\mathbf{H} \in \mathbb{R}^{m\times n}$, with $m < n$, and $\mathbf{S} = \mathbf{I}_n + \mathbf{U}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{U}$, then*

$$\kappa(\mathbf{S}) \leq 1 + \frac{\lambda_{\max}(\mathbf{B})}{\lambda_{\min}(\mathbf{R})}\lambda_{\max}(\mathbf{H}\mathbf{H}^T). \tag{13}$$

The upper bound in Equation (13) attempts to separate the influence of $\mathbf{B}$ and $\mathbf{R}$ by considering their eigenvalues separately. While this separation makes the bound easier to use for sensitivity analyses, it can also degrade the accuracy of the bound. Indeed, compared to the bound given by Equation (12), the bound given by Equation (13) does not account for any interaction between $\mathbf{B}$ and $\mathbf{R}^{-1}$, which results in a more pessimistic bound.

As an example, we consider the case where both $\mathbf{B}$ and $\mathbf{R}$ are modeled using diffusion operators as described in Section 2.3, and $\mathbf{H}$ is a selection matrix. We consider a domain of length 2000 km, composed of $n = 500$ points that are equally spaced every $h_b = 4$ km. We assume that direct observations are available at every other grid point ($m = 250$ and

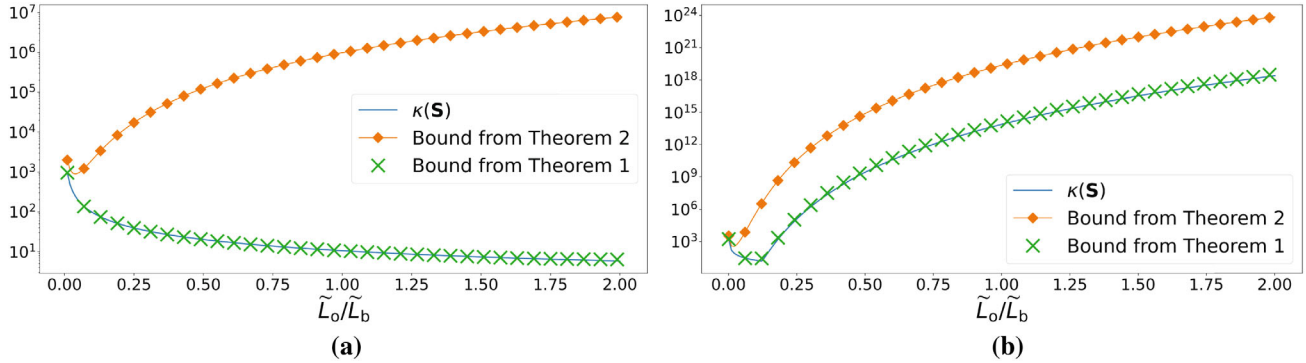



**FIGURE 1** Upper bounds on the condition number of $\mathbf{S}$ compared to the exact condition number as a function of $\widetilde{L}_o/\widetilde{L}_b = L_o h_b/L_b h_o$. (a) The correlation function of the background error is Gaussian-like ($M_b = 8$), while the correlation function of the observation error is a SOAR function ($M_o = 2$). (b) The correlation function of the background error is a SOAR function ($M_b = 2$), while the correlation function of the observation error is Gaussian-like ($M_o = 8$).

$h_o = 8$ km). We define $\widetilde{L}_b = L_b/h_b$ and $\widetilde{L}_o = L_o/h_o$ where $L_b = 60$ km is fixed and $L_o$ is allowed to vary. Figure 1 compares the two upper bounds from Theorems 1 and 2 with the exact condition number[3] for different parameter specifications in $\mathbf{B}$ and $\mathbf{R}$. Figure 1a shows the results for $M_b = 8$ and $M_o = 2$, while Figure 1b shows the results for $M_b = 2$ and $M_o = 8$, for values of $\widetilde{L}_o/\widetilde{L}_b$ ranging from 0.01 to 2. While the bound of Theorem 1 matches closely the condition number for both settings, the bound of Theorem 2 is far less accurate. This discrepancy shows that even though the diffusion models use different parameters, the structural similarities between $\mathbf{B}$ and $\mathbf{R}$ lead to a crucial interaction between them. Therefore, separating the effect of these matrices in the bound results in a pessimistic upper bound.

In this article, we are interested in obtaining explicit and accurate theoretical bounds on the condition number in terms of the key parameters of the diffusion-modeled correlation operators presented in Section 3.1. We restrict $\mathbf{H}$ to a class of uniform selection operators, which are associated with uniformly distributed observations. In Section 3.2, we study how these selection operators interact with diffusion operators. Based on the results of these two sections, we examine the conditioning of $\mathbf{S}$ in Section 3.3. Additional results are derived in Section 3.3.1 using a simplified matrix $\mathbf{S}_o$, which is equal to $\mathbf{S}$ when $\mathbf{H}$ is the identity matrix but is an approximation otherwise.

## 3.1 | Spectral properties of diffusion operators

The diffusion-modeled covariance operators for $\mathbf{B}$ and $\mathbf{R}$ in Equations (10) and (11) are defined in terms of Laplacian matrices $\mathbf{\Delta}_{h_b} \in \mathbb{R}^{n\times n}$ and $\mathbf{\Delta}_{h_o} \in \mathbb{R}^{m\times m}$, formed from a centred finite-difference discretization of the Laplacian operator on a uniform grid of resolution $h_b$ and $h_o$, respectively. On a periodic domain, $\mathbf{\Delta}_{h_b}$ and $\mathbf{\Delta}_{h_o}$ are circulant matrices[4] that are *tridiagonal* except for the first and last lines where additional non-zero elements appear in the corners due to the periodic boundary conditions. Likewise, the shifted Laplacian matrices $\mathbf{T}_b \in \mathbb{R}^{n\times n}$ and $\mathbf{T}_o \in \mathbb{R}^{m\times m}$ are circulant, near-tridiagonal matrices. Specifically, for $\mathbf{R}$, we have

$$\mathbf{\Delta}_{h_o} = \frac{1}{h_o^2}\begin{pmatrix} -2 & 1 & 0 & & 0 & 1 \\ 1 & -2 & 1 & 0 & & 0 \\ 0 & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & 0 \\ 0 & & 0 & 1 & -2 & 1 \\ 1 & 0 & & 0 & 1 & -2 \end{pmatrix} \tag{14}$$

and thus

$$\mathbf{T}_{\mathrm{o}} = \begin{pmatrix} 1+2\widetilde{L}_{\mathrm{o}}^2 & -\widetilde{L}_{\mathrm{o}}^2 & 0 & & 0 & -\widetilde{L}_{\mathrm{o}}^2 \\ -\widetilde{L}_{\mathrm{o}}^2 & 1+2\widetilde{L}_{\mathrm{o}}^2 & -\widetilde{L}_{\mathrm{o}}^2 & 0 & & 0 \\ 0 & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & 0 \\ 0 & & 0 & -\widetilde{L}_{\mathrm{o}}^2 & 1+2\widetilde{L}_{\mathrm{o}}^2 & -\widetilde{L}_{\mathrm{o}}^2 \\ -\widetilde{L}_{\mathrm{o}}^2 & 0 & & 0 & -\widetilde{L}_{\mathrm{o}}^2 & 1+2\widetilde{L}_{\mathrm{o}}^2 \end{pmatrix} \tag{15}$$

where $\widetilde{L}_{\mathrm{o}} = L_{\mathrm{o}}/h_{\mathrm{o}}$ is a non-dimensional parameter that roughly corresponds to the number of grid points over which observation-error correlations are significant. The expressions for $\mathbf{\Delta}_{h_{\mathrm{b}}}$ and $\mathbf{T}_{\mathrm{b}}$ are the same as Equations (14) and (15) with $(h_{\mathrm{b}}, \widetilde{L}_{\mathrm{b}})$ instead of $(h_{\mathrm{o}}, \widetilde{L}_{\mathrm{o}})$.

An important property of circulant matrices is that, for a given size, they all share the same eigenvectors, which form a Fourier basis.[50] Consequently, their eigenvalues can be found by taking the discrete Fourier transform of one of the rows. Let $[\mathbf{f}^{(i)}]_p$ be the $p$-th component of the $i$-th eigenvector of $\mathbf{T}_{\mathrm{o}}$ and let $\lambda_i$ be the corresponding eigenvalue:

$$\left[\mathbf{f}^{(i)}\right]_p = \frac{1}{\sqrt{m}} e^{-2\pi j \frac{ip}{m}}, \qquad \lambda_i = 1 + 4\widetilde{L}^2 \sin^2\left(\pi \frac{i}{m}\right), \; i, p \in [\![0, m-1]\!],$$

where $j$ denotes the imaginary unit ($j^2 = -1$). Since $\mathbf{T}_{\mathrm{o}}$ is symmetric, its eigenvalues are real and each of them is repeated twice; that is, $\lambda_i(\mathbf{T}_{\mathrm{o}}) = \lambda_{m-i}(\mathbf{T}_{\mathrm{o}})$, except $\lambda_0(\mathbf{T}_{\mathrm{o}})$ as the index $i$ stops at $m-1$. If $m$ is even, $\lambda_{m/2}(\mathbf{T}_{\mathrm{o}})$ is also unique.

The covariance matrices $\mathbf{B}$ and $\mathbf{R}$ in Equations (10) and (11) are proportional to a power of the inverse of $\mathbf{T}_{\mathrm{b}}$ and $\mathbf{T}_{\mathrm{o}}$, respectively. They are also circulant matrices and diagonal in a Fourier basis described by the vectors $\mathbf{f}^{(i)}$. Two circulant covariance matrices with different parameter specifications then share the same eigenvectors as long as they are applied on the same domain. In addition, since $\mathbf{T}_{\mathrm{o}}$ and $\mathbf{T}_{\mathrm{b}}$ are symmetric, the diffusion matrices are symmetric and their eigenvalues are real and proportional to powers of the inverse of the eigenvalues of $\mathbf{T}_{\mathrm{o}}$ and $\mathbf{T}_{\mathrm{b}}$. The eigenvalues of $\mathbf{B}$ and $\mathbf{R}$ are then

$$\lambda_i(\mathbf{B}) = \frac{\sigma_{\mathrm{b}}^2 \nu_{\mathrm{b}} L_{\mathrm{b}}}{h_{\mathrm{b}}} \lambda_i(\mathbf{T}_{\mathrm{b}})^{-M_{\mathrm{b}}} = \sigma_{\mathrm{b}}^2 \nu_{\mathrm{b}} \widetilde{L}_{\mathrm{b}} \left[1 + 4\widetilde{L}_{\mathrm{b}}^2 \sin^2\left(\pi \frac{i}{n}\right)\right]^{-M_{\mathrm{b}}}, \quad i \in [\![0, n-1]\!], \tag{16}$$

$$\lambda_i(\mathbf{R}) = \frac{\sigma_{\mathrm{o}}^2 \nu_{\mathrm{o}} L_{\mathrm{o}}}{h_{\mathrm{b}}} \lambda_i(\mathbf{T}_{\mathrm{o}})^{-M_{\mathrm{o}}} = \sigma_{\mathrm{o}}^2 \nu_{\mathrm{o}} \widetilde{L}_{\mathrm{o}} \left[1 + 4\widetilde{L}_{\mathrm{o}}^2 \sin^2\left(\pi \frac{i}{m}\right)\right]^{-M_{\mathrm{o}}}, \quad i \in [\![0, m-1]\!]. \tag{17}$$

## 3.2 | Influence of H as a uniform selection operator

In the previous section, we derived the eigenvalue spectra of the covariance matrices when they are defined as circulant matrices. Our simplifying assumption that the covariance and grid parameters are constant is crucial to ensure the circulant property of the covariance matrices. In this section, we exploit these results to analyse the spectrum of the matrix $\mathbf{HBH}^{\mathsf{T}}$ that appears in $\mathbf{S}$. We have already assumed that observations are evenly distributed with a grid spacing $h_{\mathrm{o}}$. We add here another assumption such that $\zeta = h_{\mathrm{o}}/h_{\mathrm{b}}$ is an integer, meaning that there are observations every $\zeta$ grid points. In this case, we refer to the resulting observation operator $\mathbf{H}$ as a *uniform selection* operator. The total number of observations is then given by $m = n/\zeta$ assuming that $\zeta$ is a divisor of $n$. While this distribution of observations is highly simplified, it allows us to analyse explicitly the sensitivity of the condition number to the correlation model parameters since it preserves to some extent the structure of $\mathbf{B}$, as described in the following lemma.

**Lemma 1.** *Let $\mathbf{B} \in \mathbb{R}^{n\times n}$ be a symmetric circulant matrix and let $\mathbf{H} \in \mathbb{R}^{n\times m}$ be a uniform selection operator where $\zeta m = n$ with $\zeta$ a positive integer. The matrix $\mathbf{HBH}^\mathsf{T} \in \mathbb{R}^{m\times m}$ is then a symmetric circulant matrix with eigenvalues*

$$\lambda_i\left(\mathbf{HBH}^\mathsf{T}\right) = \frac{1}{\zeta}\sum_{r=0}^{\zeta-1}\lambda_{i+rm}(\mathbf{B}), \tag{18}$$

*for $i \in [\![0, m-1]\!]$.*

*Proof.* Let the notation $[\mathbf{A}]_{p,q}$ denote the element on the $p$-th row and $q$-th column of any matrix $\mathbf{A}$. As $\mathbf{B}$ is a circulant matrix, it can be diagonalized in a Fourier basis:

$$\mathbf{B} = \mathbf{F}_n\mathbf{\Lambda}_\mathrm{b}\mathbf{F}_n^\mathrm{H}, \tag{19}$$

where $\mathbf{\Lambda}_\mathrm{b}$ is a diagonal matrix, and the elements of $\mathbf{F}_n$ are the (normalized) $n$-th roots of unity:

$$\forall p,q \in [\![0,n-1]\!], \quad [\mathbf{F}_n]_{p,q} = \frac{1}{\sqrt{n}}\omega_n^{pq} \quad \text{with} \quad \omega_n = e^{\frac{2\pi j}{n}}.$$

The superscript "H" stands for conjugate (Hermitian) transpose and $\mathbf{F}_n^\mathrm{H} = \mathbf{F}_n^{-1}$. Starting from Equation (19), we have

$$\mathbf{HBH}^\mathsf{T} = \mathbf{HF}_n\mathbf{\Lambda}_\mathrm{b}(\mathbf{HF}_n)^H. \tag{20}$$

The matrix $\mathbf{HF}_n$ is of dimension $m \times n$ and is composed of the rows of $\mathbf{F}_n$:

$$\forall p \in [\![0,m-1]\!], q \in [\![0,n-1]\!] \quad [\mathbf{HF}_n]_{p,q} = [\mathbf{F}_n]_{\zeta p,q} = \frac{1}{\sqrt{n}}\omega_n^{\zeta pq}.$$

The matrix $\mathbf{HF}_n$ can be linked to $\mathbf{F}_m$, the matrix that diagonalizes circulant matrices of dimension $m \times m$. The elements of the latter are the $m$-th root of unity:

$$\forall p,q \in [\![0,m-1]\!], \quad [\mathbf{F}_m]_{p,q} = \frac{1}{\sqrt{m}}\omega_m^{pq} \quad \text{with} \quad \omega_m = e^{\frac{2j\pi}{m}},$$

which, as $n = \zeta m$, can be linked to the $n$-th root of unity as $\omega_m = \omega_n^\zeta$. Consequently, for the first $m$ columns of $\mathbf{HF}_n$, we have

$$\forall p \in [\![0,m-1]\!], q \in [\![0,m-1]\!] \quad [\mathbf{HF}_n]_{p,q} = \frac{1}{\sqrt{n}}\omega_m^{pq} = \frac{1}{\sqrt{\zeta}}[\mathbf{F}_m]_{p,q}.$$

The other $n-m$ columns of $\mathbf{HF}_n$ can be characterized by using the periodicity of the $m$-th root of unity: $\omega_m^{pq} = \omega_m^{p(q+rm)}$ for any positive integer $r$. Therefore, $\mathbf{HF}_n$ is a matrix concatenated with $\zeta$ copies of $\mathbf{F}_m$:

$$\mathbf{HF}_n = \frac{1}{\sqrt{\zeta}}[\mathbf{F}_m \cdots \mathbf{F}_m].$$

Equation (20) can thus be rewritten as

$$\mathbf{HBH}^\mathsf{T} = \mathbf{F}_m\frac{1}{\sqrt{\zeta}}\left(\mathbf{I}_m \cdots \mathbf{I}_m\right)\mathbf{\Lambda}_\mathrm{b}\begin{pmatrix}\mathbf{I}_m\\ \vdots \\ \mathbf{I}_m\end{pmatrix}\frac{1}{\sqrt{\zeta}}\mathbf{F}_m^\mathrm{H} = \mathbf{F}_m\mathbf{\Lambda}'_\mathrm{b}\mathbf{F}_m^\mathrm{H},$$

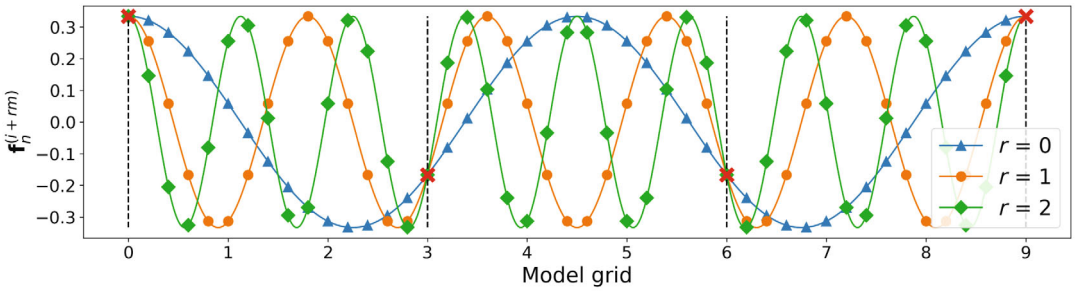



**FIGURE 2** Eigenvectors $\mathbf{f}_n^{(i+rm)}$ of a circulant matrix of size $n \times n$ with $n = 9$, for $i = 2$, $r \in [\![0, \zeta - 1]\!]$ with $\zeta = 3$ and $m = n/\zeta = 3$. On the model grid with $n$ points, these eigenvectors can be distinguished. On the observation grid with $m$ points (marked by dashed lines), they all take the same values, which are the elements of the corresponding eigenvector $\mathbf{f}_m^{(i)}$ of a circulant matrix of size $m \times m$. Note that the first and last points in the figure represent the same grid point due to the periodic boundary conditions. The vectors have been plotted with a higher resolution than the one of the model grid for the sake of clarity.

where $\mathbf{\Lambda}'_{\mathrm{b}}$ is a diagonal matrix of dimension $m \times m$. As $\mathbf{HBH}^{\mathsf{T}}$ is diagonal for the basis defined by the columns of $\mathbf{F}_m$, it is a circulant matrix. Its eigenvalues are the non-zero elements of the diagonal matrix $\mathbf{\Lambda}'_{\mathrm{b}}$, which are given by Equation (18). ∎

A matrix-vector product with $\mathbf{HBH}^{\mathsf{T}}$ is therefore in the range of the column vectors $\mathbf{f}_m^{(i)}$, weighted by the *average of $\zeta$ evenly-distributed eigenvalues* of $\mathbf{B}$. This result can be linked to the notion of *aliasing*. Different vectors from the Fourier basis of dimension $n$ (e.g., column $\mathbf{f}_n^{(i)}$) become indistinguishable and equal to the same frequency mode in the Fourier basis of dimension $m$ (e.g., column $\mathbf{f}_m^{(i)}$) once they are sub-sampled, as the highest frequencies cannot be resolved by the observation grid. This point is illustrated in Figure 2, which shows multiple distinct columns of $\mathbf{F}_n$ that all take the same values on the observation grid, which are the values of a column of $\mathbf{F}_m$. As shown in Equation (18), the weight associated with a frequency mode $\mathbf{f}_m^{(i)}$ in $\mathbf{HBH}^{\mathsf{T}}$ is the average of the weights associated with the frequency modes $\mathbf{f}_n^{(i)}$, $\mathbf{f}_n^{(i+m)}$, … , $\mathbf{f}_n^{(i+(\zeta-1)m)}$ in $\mathbf{B}$, which become equal to $\mathbf{f}_m^{(i)}$ once sub-sampled by $\mathbf{H}$.

## 3.3 | Spectral properties of the B-preconditioned Hessian matrix

Lemma 1 implies that $\mathbf{HBH}^{\mathsf{T}}$ shares the same eigenvectors as any circulant matrix of size $m \times m$, including the matrices associated with the constant-parameter diffusion operator applied on the uniform observation grid. This allows us to derive a new expression for the spectrum of $\mathbf{S}$.

**Lemma 2.** *Let $\mathbf{B} = \mathbf{U}\mathbf{U}^{\mathsf{T}} \in \mathbb{R}^{n\times n}$ and $\mathbf{R} \in \mathbb{R}^{m\times m}$, with $m < n$, be circulant matrices and let $\mathbf{H} \in \mathbb{R}^{n\times m}$ be a uniform selection operator. If $\mathbf{S} = \mathbf{I}_n + \mathbf{U}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{U}$ then the i-th eigenvalue of $\mathbf{S}$ is*

$$\lambda_i(\mathbf{S}) = \begin{cases} 1 + \frac{\lambda_i(\mathbf{HBH}^{\mathsf{T}})}{\lambda_i(\mathbf{R})} & \text{if } i \in [\![0, m-1]\!], \\ 1 & \text{otherwise.} \end{cases}$$

*Proof.* Let $\Lambda^*(\cdot)$ denote the spectrum of a matrix without its zero eigenvalues. For any matrix $\mathbf{P}$ and $\mathbf{Q}$ of respective sizes $n \times m$ and $m \times n$, with $n > m$, we know that [Theorem 21.10.1, Reference 51]

$$\Lambda^*(\mathbf{PQ}) = \Lambda^*(\mathbf{QP}). \tag{21}$$

With $\mathbf{P} = \mathbf{U}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}$ and $\mathbf{Q} = \mathbf{R}^{-1}\mathbf{H}\mathbf{U}$, Equation (21) implies that $\mathbf{U}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{U}$ has at least $n - m$ eigenvalues equal to zero, and that

$$\Lambda^*\left(\mathbf{U}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{U}\right) = \Lambda^*\left(\mathbf{R}^{-1}\mathbf{H}\mathbf{B}\mathbf{H}^{\mathsf{T}}\right).$$

Therefore, $\mathbf{S}$ has an eigenvalue of 1 with multiplicity of $n - m$ and the remaining $m$ eigenvalues are the elements of $1 + \Lambda^*\left(\mathbf{R}^{-1}\mathbf{H}\mathbf{B}\mathbf{H}^{\mathsf{T}}\right)$. Since $\mathbf{HBH}^{\mathsf{T}}$ and $\mathbf{R}^{-1}$ are circulant matrices (see Lemma 1), they are

diagonalizable in the same basis. Therefore, the eigenvalues of their product is the product of their respective eigenvalues:

$$\forall i \in [\![0, m-1]\!], \quad \lambda_i\left(\mathbf{R}^{-1}\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right) = \lambda_i\left(\mathbf{R}^{-1}\right)\lambda_i\left(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right) = \frac{\lambda_i\left(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)}{\lambda_i\left(\mathbf{R}\right)}.$$

■

We can now write the eigenvalues of $\mathbf{S}$ in terms of the constant parameters of the *diffusion-modeled covariance matrices* by using the results of Lemmas 1 and 2, and the expressions for the eigenvalues of $\mathbf{B}$ and $\mathbf{R}$ given by Equations (16) and (17), respectively.

**Theorem 3.** *Let* $\mathbf{B} = \mathbf{U}\mathbf{U}^{\mathrm{T}} \in \mathbb{R}^{n\times n}$ *and* $\mathbf{R} \in \mathbb{R}^{m\times m}$ *be circulant matrices defined by Equations* (10) *and* (11) *where the shifted Laplacian matrices* $\mathbf{T}_{\mathrm{b}}$ *and* $\mathbf{T}_{\mathrm{o}}$ *are defined in Section* 3.1. *Let* $\mathbf{H} \in \mathbb{R}^{n\times m}$ *be a uniform selection operator where* $\zeta m = n$ *with* $\zeta$ *a positive integer. If* $\mathbf{S} = \mathbf{I}_n + \mathbf{U}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{U}$ *then*

$$\lambda_i(\mathbf{S}) = \begin{cases} 1 + \alpha \sum_{r=0}^{\zeta-1} \dfrac{\left[1 + 4\widetilde{L}_{\mathrm{o}}^2 \sin^2\left(\pi \frac{i}{m}\right)\right]^{M_{\mathrm{o}}}}{\left[1 + 4\widetilde{L}_{\mathrm{b}}^2 \sin^2\left(\pi \frac{i+rm}{\zeta m}\right)\right]^{M_{\mathrm{b}}}} & \text{if } i \in [\![0, m-1]\!], \\ 1 & \text{otherwise,} \end{cases} \tag{22}$$

*with* $\widetilde{L}_{\mathrm{o}} = L_{\mathrm{o}}/h_{\mathrm{o}}$, $\widetilde{L}_{\mathrm{b}} = L_{\mathrm{b}}/h_{\mathrm{b}}$ *and*

$$\alpha = \frac{\sigma_{\mathrm{b}}^2 \nu_{\mathrm{b}} L_{\mathrm{b}}}{\sigma_{\mathrm{o}}^2 \nu_{\mathrm{o}} L_{\mathrm{o}}}.$$

*Proof.* Lemma 2 provides an expression for the eigenvalues of $\mathbf{S}$ in terms of the eigenvalues of $\mathbf{R}$ and $\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}$. The eigenvalues of $\mathbf{R}$ are known from Equation (17), and the eigenvalues of $\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}$ are obtained by applying the result of Lemma 1 to the eigenvalues of $\mathbf{B}$ from Equation (16):

$$\lambda_i\left(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right) = \frac{\sigma_{\mathrm{b}}^2 \nu_{\mathrm{b}} L_{\mathrm{b}}}{h_{\mathrm{o}}} \sum_{r=0}^{\zeta-1} \left[1 + 4\widetilde{L}_{\mathrm{b}}^2 \sin^2\left(\pi \frac{i + rm}{\zeta m}\right)\right]^{-M_{\mathrm{b}}}, \quad i \in [\![0, m-1]\!],$$

as $h_{\mathrm{o}} = \zeta h_{\mathrm{b}}$ and $n = \zeta m$. ■

From Theorem 3, it is clear that the minimum eigenvalue of $\mathbf{S}$, $\lambda_{\min}(\mathbf{S})$, is equal to one when $m < n$ (fewer observations than background variables), and is bounded below by 1 when $m = n$. The condition number of $\mathbf{S}$ is thus bounded above by the maximum eigenvalue of $\mathbf{S}$, $\lambda_{\max}(\mathbf{S})$. There is no simple analytical expression for $\lambda_{\max}(\mathbf{S})$ that can be deduced from Theorem 3. However, we can already notice that $\lambda_{\max}(\mathbf{S})$ increases with increasing ratio between the background- and observation-error variances, $\sigma_{\mathrm{b}}^2/\sigma_{\mathrm{o}}^2$. This basic dependency of the condition number on the relative variances is well known.[28,30,52]

### 3.3.1 | Spectral properties of the simplified $\mathbf{B}$-preconditioned Hessian matrix

Analysing the sensitivity of $\lambda_{\max}(\mathbf{S})$ with respect to the diffusion parameters is not straightforward from the expression given in Theorem 3. We can obtain a better understanding by approximating the effect of the matrix $\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}$ with a diffusion operator discretized directly on the observation grid. Specifically, let $\mathbf{B}_{\mathrm{o}} \in \mathbb{R}^{m\times m}$ be a diffusion operator with the same covariance parameters as $\mathbf{B} \in \mathbb{R}^{n\times n}$ but discretized on the observation grid:

$$\mathbf{B}_{\mathrm{o}} = \frac{\sigma_{\mathrm{b}}^2 \nu_{\mathrm{b}} L_{\mathrm{b}}}{h_{\mathrm{o}}} \left(\mathbf{I}_m - L_{\mathrm{b}}^2 \boldsymbol{\Delta}_{h_{\mathrm{o}}}\right)^{-M_{\mathrm{b}}}.$$

As $\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}$ and $\mathbf{B}_{\mathrm{o}}$ are two spatial discretizations of the same continuous diffusion operator, the difference between the two is solely due to the error associated with the spatial discretization. If there are direct observations at each grid point

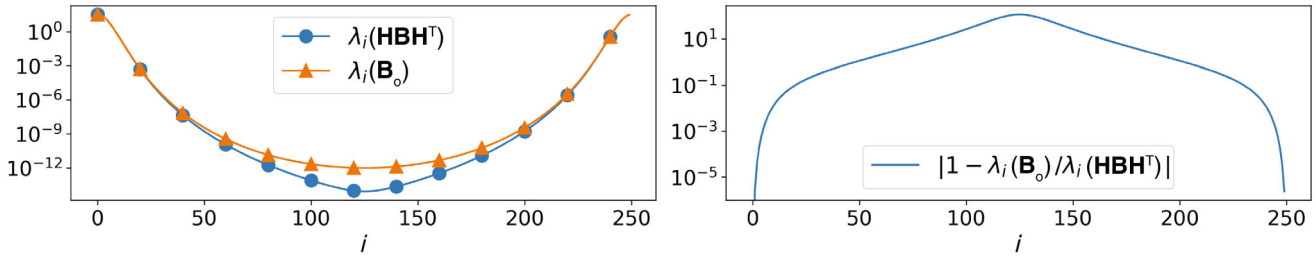



**FIGURE 3** Eigenvalues of $\mathbf{HBH}^\mathsf{T}$ and $\mathbf{B}_o$ (left panel) and their relative difference (right panel), computed for a domain of 2000 km, with a model grid of $n = 500$ points and an observation at every other model grid point ($m = 250$). The correlation model represented by $\mathbf{B}$ is an AR function of order $M_b = 8$ with a Daley length-scale of $D_b = L_b\sqrt{2M_b - 3} = 100$ km.

($\mathbf{H} = \mathbf{I}_n$), both $\mathbf{B}_o$ and $\mathbf{HBH}^\mathsf{T}$ are equal to $\mathbf{B}$ and there is no approximation. If there are less observations than grid points, $\mathbf{B}_o$ and $\mathbf{HBH}^\mathsf{T}$ still share the same eigenvectors as they are both circulant. However, they have slightly different eigenvalues due to the different spatial discretizations. As $\mathbf{B}_o$ is a diffusion operator, its eigenvalues can be deduced from the results of Section 3.1:

$$\lambda_i(\mathbf{B}_o) = \sigma_b^2 \nu_b \widetilde{L}_{b/o}\left[1 + 4\widetilde{L}_{b/o}^2 \sin^2\left(\pi\frac{i}{m}\right)\right]^{-M_b}, \tag{23}$$

where $\widetilde{L}_{b/o} = L_b/h_o$. The eigenvalues of $\mathbf{B}_o$ tend to overestimate the eigenvalues of $\mathbf{HBH}^\mathsf{T}$, with maximum relative error occurring for the smallest eigenvalues, as illustrated in Figure 3 for the case where $h_o/h_b = 2$.

Let us recall that $\mathbf{S}$ has an eigenvalue of 1 with multiplicity of $n - m$ and that the remaining $m$ eigenvalues are the elements of $1 + \Lambda^*\left(\mathbf{R}^{-1}\mathbf{HBH}^\mathsf{T}\right)$ (see Lemma 2). Approximating the matrix $\mathbf{HBH}^\mathsf{T}$ by $\mathbf{B}_o$, we are now interested in determining the eigenvalues of the matrix

$$\mathbf{S}_o = \mathbf{I}_m + \mathbf{R}^{-1}\mathbf{B}_o.$$

**Theorem 4.** *Let $\mathbf{B}_o \in \mathbb{R}^{m\times m}$ be the circulant matrix defined by Equation (10) with $h_b = h_o$, and let $\mathbf{R} \in \mathbb{R}^{m\times m}$ be the circulant matrix defined by Equation (11). The shifted Laplacian matrices $\mathbf{T}_b$ and $\mathbf{T}_o$ are defined in Section 3.1. If $\mathbf{S}_o = \mathbf{I}_m + \mathbf{R}^{-1}\mathbf{B}_o$ then*

$$\forall i \in [\![0, m-1]\!], \quad \lambda_i(\mathbf{S}_o) = 1 + \alpha\frac{\left[1 + 4\widetilde{L}_o^2 \sin^2\left(\pi\frac{i}{m}\right)\right]^{M_o}}{\left[1 + 4\widetilde{L}_{b/o}^2 \sin^2\left(\pi\frac{i}{m}\right)\right]^{M_b}},$$

*with*

$$\alpha = \frac{\sigma_b^2 \nu_b \widetilde{L}_{b/o}}{\sigma_o^2 \nu_o \widetilde{L}_o} = \frac{\sigma_b^2 \nu_b L_b}{\sigma_o^2 \nu_o L_o}. \tag{24}$$

*Proof.* The eigenvalues of $\mathbf{S}_o$ are given by $\lambda_i(\mathbf{S}_o) = 1 + \lambda_i\left(\mathbf{R}^{-1}\mathbf{B}_o\right)$. As $\mathbf{B}_o$ and $\mathbf{R}^{-1}$ are both circulant matrices, they share the same eigenvectors. Therefore, we have

$$\lambda_i(\mathbf{S}_o) = 1 + \frac{\lambda_i(\mathbf{B}_o)}{\lambda_i(\mathbf{R})}.$$

Replacing the eigenvalues of $\mathbf{B}_o$ and $\mathbf{R}$ by their expressions provided by Equations (23) and (17), respectively, yields the expression for the eigenvalues of $\mathbf{S}_o$. ∎

The next theorem provides a bound on the condition number of $\mathbf{S}_o$ by using the expression for the eigenvalues of $\mathbf{S}_o$.

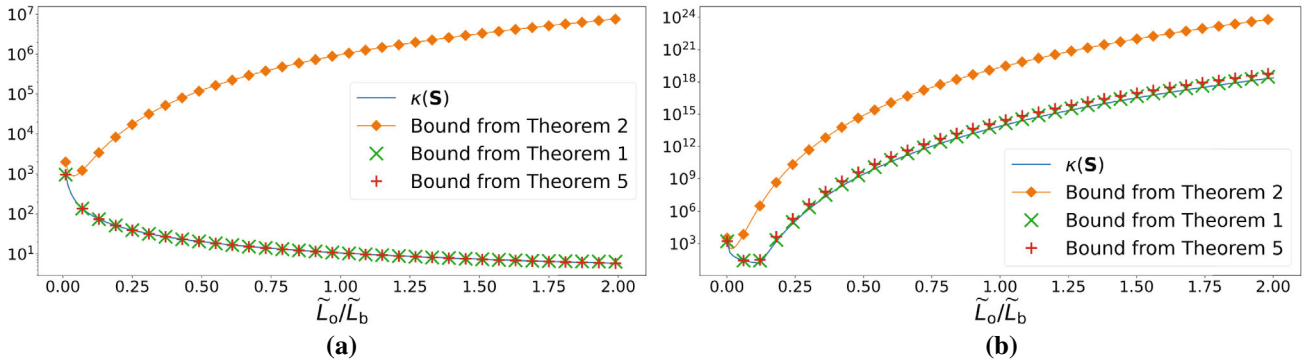



**FIGURE 4** As in Figure 1 but with an additional curve for the new bound from Theorem 5. (a) $M_o = 2, M_b = 8$. (b) $M_o = 8, M_b = 2$.

**Theorem 5.** *Let* $\mathbf{S}_o$ *be defined as in Theorem 4, and let* $\alpha$ *be given by Equation (24). Then,* $\kappa(\mathbf{S}_o) \leq \eta$ *where*

$$\eta = \begin{cases} 1 + \alpha \left(\frac{\widetilde{L}_o^2}{M_b}\right)^{M_b} \left(\frac{M_o}{\widetilde{L}_{b/o}^2}\right)^{M_o} \left(\frac{M_b - M_o}{\widetilde{L}_o^2 - \widetilde{L}_{b/o}^2}\right)^{M_b - M_o} & \text{if (i) } \widetilde{L}_o^2 M_o > \widetilde{L}_{b/o}^2 M_b; \text{(ii) } M_o < M_b; \text{ and} \\ & \text{(iii) } \widetilde{L}_{b/o}^2 M_b - \widetilde{L}_o^2 M_o > 4\widetilde{L}_{b/o}^2 \widetilde{L}_o^2 (M_o - M_b) \\ 1 + \alpha \max\left\{\frac{(1+4\widetilde{L}_o^2)^{M_o}}{(1+4\widetilde{L}_{b/o}^2)^{M_b}}, 1\right\} & \text{otherwise.} \end{cases}$$

*Proof.* See Appendix B. ∎

As explained further in this section, Theorem 5 describes the sensitivity of the condition number of $\mathbf{S}_o$ to the diffusion parameters *while keeping the bound sharp*. The sharpness of the bound is illustrated in Figure 4, where it is compared with the exact condition number of $\mathbf{S}$ and the bounds given in Theorems 1 and 2. The exact condition number has been evaluated using the extreme eigenvalues taken from the full spectrum of exact eigenvalues provided by the expression in Theorem 3. These results show that taking into account the specific structure of the covariance matrices improves the bound relative to the one given in Theorem 2.

We can further simplify the result in Theorem 5 by considering $\eta$ as a function of $\widetilde{L}_o$ only. Corollaries 1 and 2 below characterize the variations of $\eta$ with respect to $\widetilde{L}_o$ when $M_o \geq M_b$ and $M_o < M_b$, respectively.

**Corollary 1.** *Consider that* $\eta$ *defined in Theorem 5 is a function of* $\widetilde{L}_o$; $\eta = f(\widetilde{L}_o)$. *Assume that* $\widetilde{L}_o$ *has a lower bound such that* $\widetilde{L}_o\sqrt{2M_o - 1} > 1/2$ *and that condition (ii) from Theorem 5 is not met; that is,* $M_o \geq M_b$. *Then, the minimum of the function* $f$ *is unique and reached when*

$$\left(1 + 4\widetilde{L}_o^2\right)^{M_o} = \left(1 + 4\widetilde{L}_{b/o}^2\right)^{M_b}. \tag{25}$$

*Proof.* See Appendix C. ∎

**Corollary 2.** *Consider that* $\eta$ *defined in Theorem 5 is a function of* $\widetilde{L}_o$; $\eta = f(\widetilde{L}_o)$. *Assume that condition (ii) from Theorem 5 is met* ($M_o < M_b$). *Assume further that condition (iii) holds when condition (i) is satisfied. Then, the minimum of the function* $f$ *is unique and reached when*

$$L_o\sqrt{2M_o - 1} = L_b\sqrt{2M_b - 1} \tag{26}$$

*Proof.* See Appendix D. ∎

Although the assumptions in Corollaries 1 and 2 appear restrictive, they exclude cases that are of limited practical interest. In particular, values of $\widetilde{L}_o$ that are smaller than the observation grid resolution $h_o$ correspond to observation errors that are effectively uncorrelated. To avoid this case, we focus on values of $\widetilde{L}_o \geq 1$, which automatically fulfils the

condition on the lower bound in Corollary 1. Alternatively, the condition on the lower bound can be seen as restricting the Stein length-scale $\rho_o = L_o\sqrt{2M_o - 1}$ (see Equation 8 with $w = 1$) to be greater than half the grid resolution. In Corollary 2, we assume that condition (iii) of Theorem 5 holds if condition (i) is satisfied. To simplify condition (iii), we can impose a practical bound on the value of $\widetilde{L}_{b/o}$. For instance, we are not interested in cases where the length-scale $L_b$ is smaller than $h_o$, that is, $\widetilde{L}_{b/o} \leq 1$. More generally, we can assume that $\widetilde{L}_{b/o}$ is bounded below by a positive scalar $\widetilde{L}_{\min}$, which leads to the next corollary.

**Corollary 3.** *Assume that there exists a positive scalar $\widetilde{L}_{\min}$ such that*

$$\widetilde{L}_{b/o} \geq \widetilde{L}_{\min},$$

*and that condition (ii) of Theorem 5 holds. Then, condition (iii) of Theorem 5 is met if*

$$M_b \leq 2\left(1 + 4\widetilde{L}^2_{\min}\right). \tag{27}$$

*Proof.* Let us define $r_M = M_o/M_b$. Using the assumption on the length-scale, condition (iii) of Theorem 5 can be rewritten as

$$\widetilde{L}^2_{\min} + 4\widetilde{L}^2_{\min}\widetilde{L}^2_o(1 - r_M) - \widetilde{L}^2_o r_M \geq 0.$$

As $M_o$ and $M_b$ are assumed to be even integers and as condition (ii) is met ($M_o < M_b$), we know that $r_M \leq (M_b - 2)/M_b$. Using this relation, we obtain that

$$M_b\widetilde{L}^2_{\min} + 8\widetilde{L}^2_{\min}\widetilde{L}^2_o - \widetilde{L}^2_o(M_b - 2) \geq 0,$$

which can be rearranged to give

$$M_b\left(1 - \widetilde{L}^2_{\min}/\widetilde{L}^2_o\right) \leq 2\left(1 + 4\widetilde{L}^2_{\min}\right).$$

Since $\widetilde{L}^2_{\min}/\widetilde{L}^2_o$ is positive, we obtain the inequality (27). ∎

Taking $\widetilde{L}_{\min} = 1$ in Equation (27) results in $M_b \leq 10$. Increasing $M_b$ beyond 10 has little practical value as the correlation function is already approximately Gaussian with this value.

For the case where **B** and **R** are modeled with SOAR functions ($M_b = M_o = 2$),[30] point out that, for fixed $L_b$, the minimum of their upper bound for the condition number of the **B**-preconditioned Hessian matrix is found by setting $L_o = L_b$. Corollaries 1 and 2 confirm this result and extend it to other AR functions ($M_o = M_b > 2$). They also cover cases where the order of the AR functions differs between **B** and **R** ($M_b \neq M_o$), in which case the function defining the upper bound on the condition number, $\eta = f(\widetilde{L}_o)$, does not reach its minimum value when $L_o = L_b$.

If $M_o > M_b$ then $\widetilde{L}_o$ can be much smaller than $\widetilde{L}_{b/o}$ to attain the minimum of the function $f(\widetilde{L}_o)$. For example, if $M_b = 2$, $M_o = 10$, and $\widetilde{L}_o = 1.5$, then $\widetilde{L}_{b/o}$ needs to be 158 to satisfy Equation (25) of Corollary 1. The correlation functions with fixed values of $\left(1 + 4\widetilde{L}^2\right)^M$ have very different range for low values of $M$ as illustrated in Figure 5a. On the other hand, if $M_o < M_b$ then Corollary 2 states that the minimum value is attained when the Stein length-scales $\rho_o = L_o\sqrt{2M_o - 1}$ and $\rho_b = L_b\sqrt{2M_b - 1}$ are equal. Note that, unlike condition (25), condition (26) is independent of $h_o$. As shown in Figure 5b, the correlation functions with fixed values of $\rho$ are very similar for different values of $M$.

For an alternative interpretation of Corollaries 1 and 2, we can recast Equations (25) and (26) in terms of the Daley length-scales $D_o$ and $D_b$ (Equation 7), which is the length-scale parameter we will use to interpret the numerical experiments in the following sections. Assuming $M_b > 1$ and $M_o > 1$, we have

$$L_b = \frac{D_b}{\sqrt{2M_b - 3}} \quad \text{and} \quad L_o = \frac{D_o}{\sqrt{2M_o - 3}}. \tag{28}$$

If $M_o = M_b$ then the minimum is reached when $D_o = D_b$.

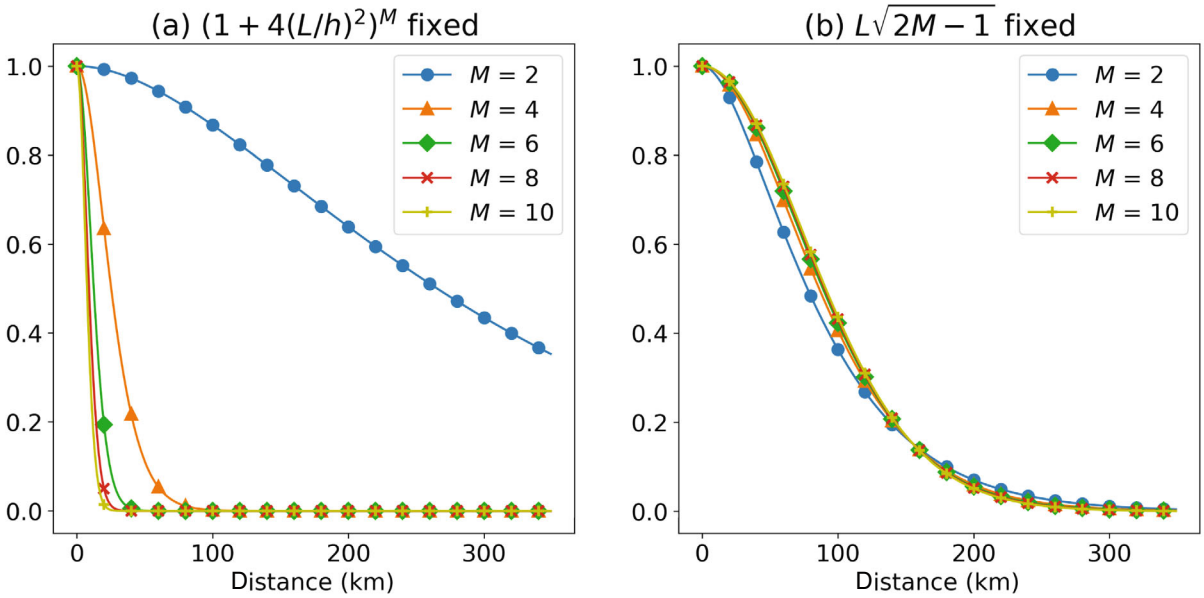



**FIGURE 5** AR correlation functions (Equation A2) displayed for different values of $M$. For each $M$, the value of $\widetilde{L}$ is chosen such that the quantities (a) $\left(1+4\widetilde{L}^2\right)^M$ and (b) $L\sqrt{2M-1}$ from Corollaries 1 and 2, respectively, are kept constant. The corresponding values of $L$, the Stein length-scale $\rho = L\sqrt{2M-1}$, and the Daley length-scale $D = L\sqrt{2M-3}$ can be found in Table 1.

**TABLE 1** Values of the length-scale parameter $L = \widetilde{L}h$ where $h = 1$ km, the Stein length-scale $\rho = L\sqrt{2M-1}$, and the Daley length-scale $D = L\sqrt{2M-3}$ associated with the curves in Figure 5.

| | $(1+4\widetilde{L}^2)^M$ fixed | | | $L\sqrt{2M-1}$ fixed | | |
|---|---|---|---|---|---|---|
| $M$ | $L$ (km) | $\rho$ (km) | $D$ (km) | $L$ (km) | $\rho$ (km) | $D$ (km) |
| 2 | 158.1 | 273.8 | 158.1 | 46.2 | 80.0 | 46.2 |
| 4 | 8.9 | 23.5 | 19.9 | 30.2 | 80.0 | 67.6 |
| 6 | 3.4 | 11.2 | 10.1 | 24.1 | 80.0 | 72.4 |
| 8 | 2.0 | 7.7 | 7.4 | 20.7 | 80.0 | 74.5 |
| 10 | 1.5 | 6.5 | 6.2 | 18.3 | 80.0 | 75.7 |

*Note*: The fixed values of $\left(1+4\widetilde{L}^2\right)^M$ and $L\sqrt{2M-1}$ are $10^{10}$ and 80 km, respectively.

If $M_o > M_b$ then Equation (25) translates as

$$\left(1+\frac{4D_o^2}{h_o^2(2M_o-3)}\right)^{M_o} = \left(1+\frac{4D_b^2}{h_o^2(2M_b-3)}\right)^{M_b}. \tag{29}$$

The location of the minima is very sensitive to $M_o$ and $M_b$ since they appear as exponents in Equation (29). While a small change of $M_o$ from 8 to 10 would have limited effect on the correlation function, it can have a drastic effect on the quantities in Equation (29). In turn, this can significantly affect the condition number (as seen from Theorem 5) as well as the criteria in Corollary 1. This property can be detrimental if ignored, but can also be exploited to our advantage to improve the conditioning without significantly altering the correlation shape, as will be illustrated in Section 4.

If $M_o < M_b$ then Equation (26) translates as

$$D_o^2\left(\frac{2M_o-1}{2M_o-3}\right) = D_b^2\left(\frac{2M_b-1}{2M_b-3}\right), \tag{30}$$

from which we can deduce that the minimum is reached when $D_o < D_b$ (cf. $L_o > L_b$ and $\rho_o = \rho_b$). This is evident from the last column of Table 1, which shows $D$ increasing with increasing $M$. The ratio between $D_b$ and $D_o$ reaches at most 1.6 for the limiting values of $M_o = 2$ and $M_b = 10$.

Equations (25) and (26) (respectively, Equations 29 and 30) provide simple criteria that can be used to adjust the value of $L_o$ (respectively, $D_o$) to minimize the condition number of the **B**-preconditioned Hessian matrix. From this perspective, we can use Corollaries 1 and 2 as the basis of a method for *reconditioning* observation-error covariance matrices that account for spatial correlations with parametric functions from the Matérn family. This would be complementary to existing methods for reconditioning sample covariance matrices, for example, to represent inter-channel error correlations in satellite observations.[53,54] For more complex problems, where the assumptions of these corollaries are not perfectly satisfied, we can still use criteria (25) and (26) (or 29 and 30) as a guideline for adjusting covariance parameters in **B** and **R** to improve the conditioning of the **B**-preconditioned Hessian matrix.

### 3.3.2 | Condition number estimates with correlated and uncorrelated observation errors

In this section, we compare the condition number of **S** for different values of the correlation parameter pairs $(M_o, D_o)$ and $(M_b, D_b)$. The condition number $\kappa(\mathbf{S})$ is computed using the (exact) analytical expression of the eigenvalues of **S** given in Theorem 3. In addition, we compute the exact 'optimal' parameter pairs (i.e., those that minimize the condition number) and compare them with those predicted by the optimality criteria in Corollaries 1 and 2. As this theorem applies to the matrix $\mathbf{S}_o$, and not **S**, these optimality criteria are only exact when there is a direct observation at each grid point.

In presenting the results, we choose to normalize $\kappa(\mathbf{S})$ by $\kappa(\mathbf{S}_u)$ where $\mathbf{S}_u$ is given by Equation (6) with $\mathbf{R} = \sigma_o^2 \mathbf{I}_m$; that is, with observation-error correlations neglected. An analytical expression for the eigenvalues of $\mathbf{S}_u$ can be derived directly from Equation (22) of Theorem 3 by setting $M_o = 0$ (no diffusion) and $\nu_o \widetilde{L}_o = \gamma^2 / h_o = 1$ (exact normalization):

$$\forall i \in [\![0, n-1]\!], \quad \lambda_i(\mathbf{S}_u) = \begin{cases} 1 + \alpha_u \sum_{r=0}^{\zeta-1} \left[ 1 + 4\widetilde{L}_b^2 \sin^2\left( \pi \frac{i+rm}{\zeta m} \right) \right]^{-M_b} & \text{if } i \in [\![0, m-1]\!], \\ 1 & \text{otherwise,} \end{cases}$$

where

$$\alpha_u = \frac{\sigma_b^2 \nu_b L_b}{\sigma_o^2 h_o}.$$

As we are considering the case where there are fewer observations than grid points ($n > m$), the minimum eigenvalue of $\mathbf{S}_u$ is one. The condition number of $\mathbf{S}_u$ is thus computed as

$$\kappa(\mathbf{S}_u) = 1 + \alpha_u \max_{i \in [\![0, m-1]\!]} \sum_{r=0}^{\zeta-1} \left[ 1 + 4\widetilde{L}_b^2 \sin^2\left( \pi \frac{i+rm}{\zeta m} \right) \right]^{-M_b}. \tag{31}$$

Note that the maximum of the sum in Equation (31) is larger than one (as it is larger than one at least for $i = 0$) and approximately equal to one for parameter values of interest; that is, for $M_b \geq 2$ and $\widetilde{L}_b \geq 1$, its maximum is less than 1.04. The condition number of $\mathbf{S}_u$ is thus dominated by $\alpha_u$.

We denote $\chi$ the ratio of condition numbers:

$$\chi = \frac{\kappa(\mathbf{S})}{\kappa(\mathbf{S}_u)}. \tag{32}$$

As $M_o$ and $D_o$ have no effect on $\kappa(\mathbf{S}_u)$, variations of $\chi$ with respect to these parameters will reflect variations of $\kappa(\mathbf{S})$. If $\chi < 1$ then accounting for correlated observation error will improve the conditioning of **S** and thus we can expect the convergence rate of CG to be improved. Conversely, if $\chi > 1$ then accounting for correlated observation error will degrade the conditioning of **S** and we can expect the convergence rate of CG to be degraded.

In the following, we will compute the condition numbers as a function of the Daley length-scales defined in Equation (28). Furthermore, since we are mainly interested in the sensitivity of the condition number to the correlation model parameters, we will assume that the background- and observation-error variances are equal ($\sigma_b^2 / \sigma_o^2 = 1$). We

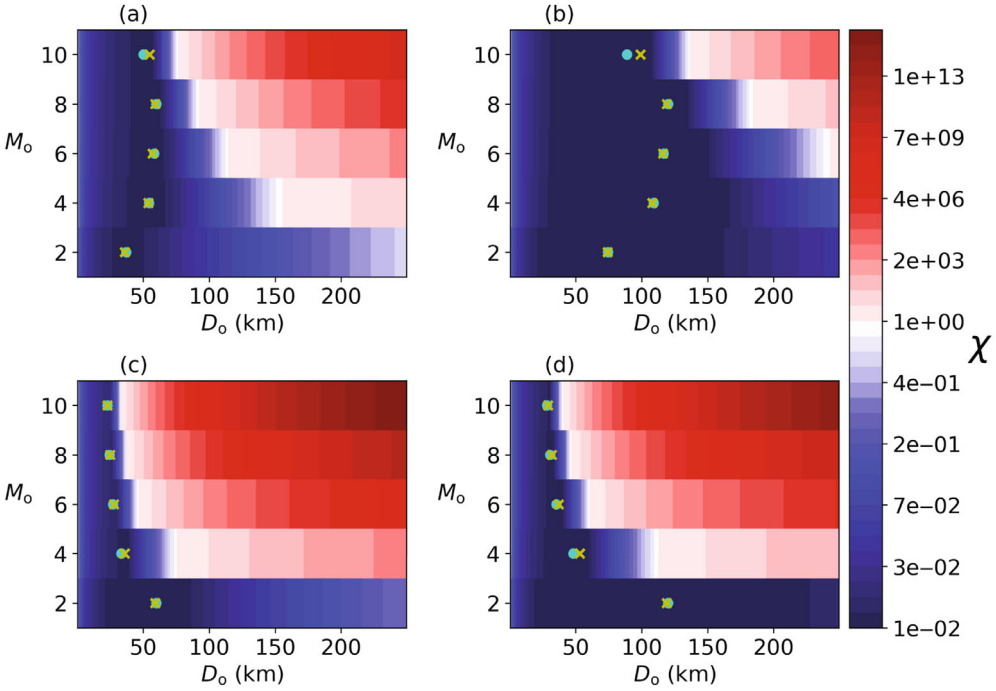



**FIGURE 6** The ratio $\chi$ (Equation 32) is plotted for a fixed parameter pair $(M_b, D_b)$ per panel (indicated in the title), and values of $M_o$ and $D_o$ that vary along the axes. The cyan circles mark the minima predicted by Corollaries 1 and 2. The yellow crosses mark the true minima. Note that the color palette uses a logarithmic scale with a different range below and above $\chi = 1$. (a) $D_b = 60$ km; $M_b = 8$. (b) $D_b = 120$ km; $M_b = 8$. (c) $D_b = 60$ km; $M_b = 2$. (d) $D_b = 120$ km; $M_b = 2$.

consider a domain of length 2000 km, composed of $n = 500$ points that are equally spaced every $h_b = 4$ km. We assume that a direct observation is available every other grid point ($\zeta = 2$, $m = 250$, $h_o = 8$ km).

Figure 6 shows $\chi$ as a function of $D_o$ (abscissa) and $M_o$ (ordinate) for different parameter pairs $(M_b, D_b)$ indicated in the title of each panel. The zones in blue (respectively, red) correspond to parameter pairs $(M_o, D_o)$ that improve (respectively, degrade) the condition number. When $M_o \leq M_b$ and $D_o \leq D_b$, the conditioning is systematically improved. An improvement is also possible when $D_o \geq D_b$ if $M_o$ is small enough. However, when $M_o$ becomes too large compared to $M_b$ or when $D_o$ becomes too large compared to $D_b$, the conditioning is degraded and can become significantly degraded even for modest changes in the parameter values. For example, when $M_b = 8$ and $D_b = 60$ km (Figure 6a), and $M_o = 10$, $\chi$ (and thus $\kappa(\mathbf{S})$) increases by several orders of magnitude when the value of $D_o$ is increased to less than double $D_b$. When $D_o$ is approximately four times $D_b$, $\chi$ reaches $10^{10}$ (top right corner of Figure 6a). In these cases, we can expect the convergence of CG to be significantly affected, as illustrated later in Section 4.

As predicted by Corollaries 1 and 2, if $D_b$, $M_b$, and $M_o$ are fixed, then $\kappa(\mathbf{S})$ admits a unique minimum. When $M_o > M_b$, the minima predicted by Corollary 1 are visibly shifted towards lower values of $D_o$ (cf. circles and crosses in Figure 6). This shift corresponds to an increase of the condition number of up to 5%. As the variations of the condition number studied here cover a range of more than 10 orders of magnitude, this increase is acceptable. When $M_o < M_b$, there is no significant difference in the position of the minima predicted by Corollary 2 and the exact minima; there is an increase of the condition number between the predicted minima and exact minima that is smaller than 0.1%. The pattern is similar with each fixed settings for $(M_b, D_b)$ (i.e., each panel of Figure 6). If $D_b$ increases (decreases) then the 'optimal' values of $D_o$ are shifted to the right (left) towards larger (smaller) values of $D_o$ (cf. Figure 6a–d).

## 4 | NUMERICAL EXPERIMENTS

In this section, we illustrate how different covariance parameter settings influence the performance of the CG minimization. We evaluate the convergence rate in relation to the condition number diagnostic $\chi$ presented in Section 3.3.2 (see Figure 6) and the results of Corollaries 1 and 2.

## 4.1 | Experimental framework

As in Section 3.3.2, we define our baseline 1D-Var experiment as one in which the background- and observation-error variances are taken to be equal, with their actual values set to one unit ($\sigma_b^2 = \sigma_o^2 = 1$). The domain is periodic with length 2000 km and there are $n = 500$ grid points ($h_b = 4$ km). We define $\mathbf{H}$ as a selection operator where direct observations are assumed to be available at every other grid point ($m = 250$, $h_o = 8$ km).

We consider different 'scenarios' where observations with perfectly known error correlations are assimilated together with a background state that also has perfectly known error correlations. We start by defining a 'true state', $\mathbf{x}_t$, which is specified by an analytical function. As $\mathbf{H}$ is linear in our framework, the actual choice of the true state has no impact on the performance of the CG minimization as it is subtracts out from the innovation vector. The background state and observations are then generated by adding to the true state, unbiased random perturbations of covariance matrices $\mathbf{B}$ and $\mathbf{R}$, respectively. Specifically, let $\hat{\epsilon}_b$ and $\hat{\epsilon}_o$ be normally-distributed vectors with zero mean and covariance matrix equal to the identity matrix. We can generate many realizations of $\hat{\epsilon}_b \sim N(\mathbf{0}, \mathbf{I}_n)$ and $\hat{\epsilon}_o \sim N(\mathbf{0}, \mathbf{I}_m)$ using a random number generator. Then, using the factored covariance matrices $\mathbf{B} = \mathbf{U}\mathbf{U}^T$ and $\mathbf{R} = \mathbf{V}\mathbf{V}^T$, we define

$$\begin{aligned} \mathbf{x}_b &= \mathbf{x}_t + \boldsymbol{\epsilon}_b, \\ \mathbf{y}_o &= \mathbf{H}\mathbf{x}_t + \boldsymbol{\epsilon}_o, \end{aligned} \tag{33}$$

where $\boldsymbol{\epsilon}_b = \mathbf{U}\hat{\epsilon}_b$ and $\boldsymbol{\epsilon}_o = \mathbf{V}\hat{\epsilon}_o$. By construction, $\mathbb{E}[\boldsymbol{\epsilon}_b\boldsymbol{\epsilon}_b^T] = \mathbf{B}$ and $\mathbb{E}[\boldsymbol{\epsilon}_o\boldsymbol{\epsilon}_o^T] = \mathbf{R}$ where $\mathbb{E}[\ ]$ is the expectation operator.

To assess the convergence rate of the CG algorithm at each iteration, it is common to monitor the reduction of the cost function or, equivalently, the reduction of the $\mathbf{A}$-norm of the analysis (solution) error. However, if we want to compare the convergence rate of CG with different $\mathbf{R}$, the $\mathbf{A}$-norm is not appropriate since it depends on $\mathbf{R}$ and thus does not represent the same quantity in all cases. Since we are working with an idealized system for which the true state $\mathbf{x}_t$ is known, we have access to alternative metrics that would not be available in a realistic system.

At the $\ell$-th iteration of the CG algorithm, an increment $\delta\mathbf{x}_\ell$ is produced. We can deduce from this increment the analysis error that would result if the CG algorithm was stopped at the $\ell$-th iteration:

$$\boldsymbol{\epsilon}_a^{(\ell)} = \mathbf{x}_b + \delta\mathbf{x}_\ell - \mathbf{x}_t.$$

In each experiment, there is a random component in the generation of the background and observations, which will affect $\boldsymbol{\epsilon}_a^{(\ell)}$. By performing multiple experiments with different right-hand sides ($\mathbf{b}$ in Equation 5), we can obtain multiple realizations of $\boldsymbol{\epsilon}_a^{(\ell)}$ from which analysis-error statistics can be deduced. In particular, we can estimate at each iteration the total analysis-error variance or, equivalently, the trace of the analysis-error covariance matrix. This is the quantity that is minimized explicitly in a statistical analysis based on the best linear unbiased estimator (BLUE). It is well known that, when the constraints are linear and when the background and observation errors are normally distributed, the minimizing solution of the cost function of variational data assimilation is equivalent to the BLUE when both are formulated under the same assumptions.[55]

As a diagnostic, we compute the square root of the average of the total analysis-error variance:

$$\sigma_a^{(\ell)} = \sqrt{\frac{1}{n}\mathbb{E}\left[\mathrm{Tr}\left((\boldsymbol{\epsilon}_a^{(\ell)})(\boldsymbol{\epsilon}_a^{(\ell)})^T\right)\right]} = \sqrt{\frac{1}{n}\mathbb{E}\left[(\boldsymbol{\epsilon}_a^{(\ell)})^T(\boldsymbol{\epsilon}_a^{(\ell)})\right]}, \tag{34}$$

where $\mathbb{E}$ denotes the expectation operator and Tr denotes the trace operator. We approximate the expectation operator as an average of 1000 realizations with random right-hand sides.

This metric can be used to assess not only the convergence rate of the minimization on which we focused in the previous sections, but also the accuracy of the solution at each iteration. We expect the solution of the minimization at full convergence to be more accurate when the actual observation-error correlations are accounted for. If the condition number is reduced by a non-diagonal $\mathbf{R}$ (i.e., $\chi < 1$ as in the 'blue zone' of Figure 6) then the minimization should converge faster. In this situation, we can expect the solution to be more accurate no matter when the minimization is stopped. On the other hand, if the condition number is increased by a non-diagonal $\mathbf{R}$ (i.e., $\chi > 1$ as in the 'red zone' of Figure 6) then we can expect the convergence rate to be slower. In this situation, it is not clear whether a non-diagonal $\mathbf{R}$ is beneficial to the analysis or not, as there is a trade-off between the convergence rate and the expected accuracy at full convergence. Monitoring the analysis error at each iteration allows us to visualize this trade-off as it indicates, at each iteration, how accurate the analysis would be (on average) if the convergence was stopped at this point.

A natural choice of normalization for $\sigma_a^{(\ell)}$ is its initial value $\sigma_a^{(0)}$, which is equal to $\sigma_b$ in the experiments as the initial increment $\delta\mathbf{x}_0$ is zero. The quantity $\sigma_a^{(\ell)}/\sigma_a^{(0)}$ thus indicates the relative error reduction on each iteration of CG. We denote $\sigma_a^*$ the value of $\sigma_a^{(\ell)}$ at full convergence of CG. If the specifications of $\mathbf{B}$ and $\mathbf{R}$ used to compute the analysis match the actual error statistics, this quantity should become equal to its theoretical minimum, $\sigma_a^{\text{opt}}$, which can be computed directly from the trace of the theoretical analysis-error covariance matrix:

$$\sigma_a^{\text{opt}} = \sqrt{\frac{1}{n}\text{Tr}\left[\left(\mathbf{B}^{-1} + \mathbf{H}^{\text{T}}\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\right]}, \tag{35}$$

where $\mathbf{B}$ and $\mathbf{R}$ are the same as those used to generate the random errors in Equation (33).

In the experiments, $\mathbf{R}$ denotes the 'true' observation-error covariance matrix used to generate the spatially-correlated random errors that are added to the observations. The matrix $\widetilde{\mathbf{R}}_1 = \sigma_o^2\mathbf{I}_m$ is a diagonal approximation where $\sigma_o^2$ is the same constant variance used in the 'true' $\mathbf{R}$. This corresponds to the common case where spatially-correlated observation errors are ignored in the weighting matrix in the cost function, which is inconsistent with the statistical properties of the observations that are assimilated. The third scenario also uses a diagonal matrix, $\widetilde{\mathbf{R}}_2 = \upsilon\,\sigma_o^2\mathbf{I}_m$, but the variances are multiplied by an inflation factor ($\upsilon$) to mitigate the effect of neglecting the error correlations. This procedure is common in real-data assimilation problems, to avoid overfitting observations at large spatial scales while retaining a simple covariance matrix. In practice, the inflation factor is usually estimated empirically. In our experiments, we can determine the best-possible inflation factor by minimizing $\sigma_a^*$ with respect to $\upsilon$. As $\sigma_a^*$ behaves approximately as a convex function of $\upsilon$, this can be achieved by computing $\sigma_a^*$ for increasing values of $\upsilon$ until it stops decreasing (i.e., until the observations are no longer overfit). This method cannot be applied in an operational context as it requires access to the true state. Even with a performance metric that uses a proxy for the true state, the cost of the procedure would be prohibitive as thousands of realizations of $\sigma_a^*$ are required. The experiments using $\widetilde{\mathbf{R}}_2$ thus represent the best inflation can offer rather than what could be achieved in practice.

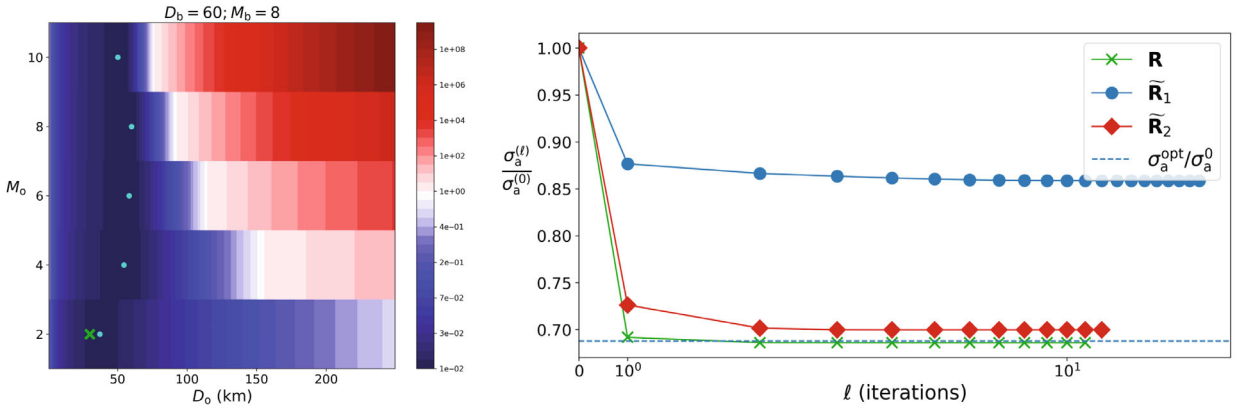
## 4.2 | Results

In the first set of experiments, we consider the case where the background- and observation-error correlation parameters are in the regime $M_o < M_b$ and $D_o < D_b$. The observation-error correlation parameters are set to $M_o = 2$ and $D_o = 30$ km, which corresponds to a SOAR function as used in Reference 30. These values are roughly similar to those proposed by Reference 18, where the parameter settings were determined to provide a suitable fit of a diffusion-model to error correlation estimates of certain satellite radiance observations.[7] The background-error correlation parameters are set to $M_b = 8$ and $D_b = 60$ km, which makes the correlation function more Gaussian-like than that of $\mathbf{R}$. The correlation length-scale of $\mathbf{B}$ is double the correlation length-scale of $\mathbf{R}$. These are the same $\mathbf{B}$ parameters that were used in Figure 6. With these parameters, we know that $\chi < 1$ (Figure 7a), which means that the condition number is reduced when observation-error correlations are accounted for.

The average error-convergence curves from the 1D-Var experiments with $\mathbf{R}$, $\widetilde{\mathbf{R}}_1$, and $\widetilde{\mathbf{R}}_2$ are shown in Figure 7b. For each experiment, minimizations are performed in parallel for all 1000 realizations of the random right-hand side and are stopped when the 2-norm of the residual normalized by its initial value reaches $10^{-6}$. Convergence is achieved rather quickly, taking about 20 iterations with $\widetilde{\mathbf{R}}_1$ and about 10 iterations with $\mathbf{R}$ and $\widetilde{\mathbf{R}}_2$.

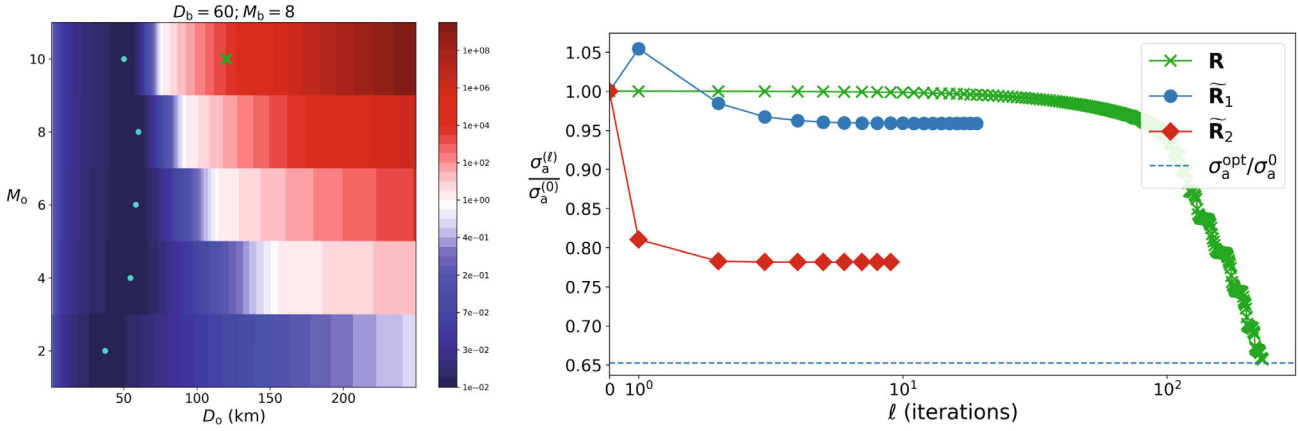
For the experiment with $\widetilde{\mathbf{R}}_1$, the analysis-error standard deviation is only reduced by about 15% at full convergence, compared to the theoretical limit of 32%. In this set-up, the optimal variance inflation factor is approximately equal to 10.5. Inflating the error variances significantly improves the error reduction (30%). However, the theoretical minimum error cannot be reached, even with an inflation factor that has been optimized for this specific experiment. It is also important to remark that the experiment with inflated variances converges faster than the experiment with the original variances. This is consistent with Equation (31), which shows that, for large $\alpha_u$, the condition number of $\mathbf{S}_u$ is approximately inversely proportional to the observation-error variance and is thus divided by 10.5 in this case. Best results are obtained with the true $\mathbf{R}$. First, the convergence rate is the fastest of the three experiments. Second, on each iteration, the solution is more accurate than the solutions from either $\widetilde{\mathbf{R}}_1$ or $\widetilde{\mathbf{R}}_2$. At full convergence, the solution attains the theoretical minimum error.

We now consider a case where the parameter values of the observation-error correlations suggest that the convergence rate of the minimization will be degraded (i.e., $\chi > 1$). Using Figure 6, we can select parameter values that will increase the condition number. In particular, we set $M_o = 10$ and $D_o = 120$ km, while keeping the background-error parameter values




**(a)** $\chi$ for different covariance parameters **(b)** CG minimisation

**FIGURE 7** (a) Same as Figure 6 but with a cross sign added to indicate the parameter pair $(M_o, D_o)$ used for the 1D-Var experiments in panel (b). (b) $\sigma_a^{(\ell)}/\sigma_a^{(0)}$ (Equation 34) as a function of CG iteration count $\ell$ for three 1D-Var experiments with the same covariance parameters for $\mathbf{B}$ ($\sigma_b^2 = 1$, $M_b = 8$, $D_b = 60$ km) but different covariance parameters for $\mathbf{R}$: (1) $\mathbf{R}$ with the 'true' correlation parameters ($\sigma_o^2 = 1$, $M_o = 2$, $D_o = 30$ km); (2) a diagonal approximation, $\widetilde{\mathbf{R}}_1 = \sigma_o^2 \mathbf{I}_m$ with $\sigma_o^2 = 1$; (3) a diagonal approximation with inflated variances, $\widetilde{\mathbf{R}}_2 = \upsilon\, \sigma_o^2 \mathbf{I}_m$ where $\upsilon = 10.5$ is an optimally-estimated inflation factor. The theoretical minimum analysis-error ratio $\sigma_a^{opt}/\sigma_a^{(0)}$ (Equation 35) is marked by a horizontal dashed line.



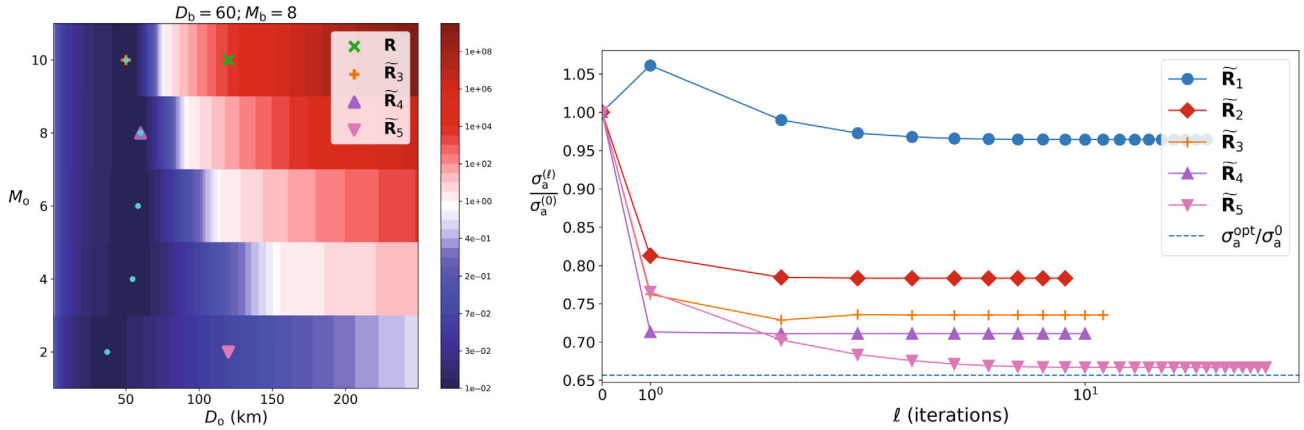

**(a)** $\chi$ for different covariance parameters **(b)** CG minimisation

**FIGURE 8** Same as Figure 7 but with a different parameter pair ($M_o = 10, D_o = 120$ km), indicated by the cross sign in panel (a). The optimal inflation factor in panel (b) is $\upsilon = 17$.

unchanged. In this set-up, the observation-error length-scale $D_o$ is double the background-error correlation length-scale. With this set of parameter values, the condition number is increased by a factor of $10^4$. In this set-up, the theoretical minimum error $\sigma_a^{opt}/\sigma_b$ is lower than in the previous experiment: 65% instead of 68%. This decrease means that observations with highly correlated errors 'complement' the background better than those of the previous experiment.

As shown in Figure 8b, while the minimization with $\mathbf{R}$ does reach the theoretical minimum, it requires about 200 iterations to converge. If the minimization was terminated in its early iterations ($< 50$) then the analysis would be hardly better than that of the background and not as accurate as the solutions from either the $\widetilde{\mathbf{R}}_1$ or $\widetilde{\mathbf{R}}_2$ experiments. In this case, it would be clearly detrimental to account for the observation-error correlations instead of ignoring them.

It is interesting to note that the convergence curve for $\widetilde{\mathbf{R}}_1$ in Figure 8b differs from the one in Figure 7b even though the Hessian matrix $\mathbf{S}$ for this case is the same in both experiments. On the other hand, the assimilated observations are different in each experiment as they have different correlated errors. The difference in the convergence curves in the two experiments thus highlights the role of the right-hand side (which depends on the observations) of the system on the convergence rate of CG. While this is an important issue, we have not attempted to address it in this article.




**(a)** $\chi$ for different covariance parameters **(b)** CG minimisation

**FIGURE 9** Same as Figure 8 but with different parameter pairs ($M_o = 10$, $D_o = 50$ km), ($M_o = 8$, $D_o = 60$ km), and ($M_o = 2$, $D_o = 120$ km) for $\mathbf{R}$ in panel (b) (experiments labeled $\widetilde{\mathbf{R}}_3$, $\widetilde{\mathbf{R}}_4$, and $\widetilde{\mathbf{R}}_5$, respectively) compared to the parameter pair ($M_o = 10$, $D_o = 120$ km) used to generate the observations. The different pairs are indicated by the different colored symbols in panel (a). The optimal inflation factor in panel (b) is $\upsilon = 17$.

The experiment $\widetilde{\mathbf{R}}_1$ results in an error reduction of only 5%, compared to the theoretical minimum of 35%. Moreover, the error reduction is non-monotonic, which is symptomatic of a more concerning issue: as $\widetilde{\mathbf{R}}_1$ is an approximation of the actual error covariances, there is no guarantee that the analysis will be more accurate than the background (even at full convergence). In this case, due to the neglected correlations in $\widetilde{\mathbf{R}}_1$, the contribution of the observations to the analysis is larger than it should be. This has a more detrimental impact on the analysis when the observations are less accurate than the background: repeating the experiment with larger observation error results in a monotonically increasing analysis error (not shown).

In the current scenario, the optimal variance inflation factor is approximately equal to 17, and the experiment with $\widetilde{\mathbf{R}}_2$ gives the best results when using a modest number of iterations ($< 150$). It produces a similar, rapid convergence rate as in the previous scenario (Figure 7b) and produces an accurate analysis, with a 23% error reduction compared to the theoretical minimum error reduction of 35%.

Rather than adopting a diagonal approximation, an alternative approach would be to adjust the correlation parameters to accelerate the convergence rate while trying to retain the salient features of the original correlation function, which in principle should correspond to our best available estimate of the actual correlation function. As discussed in Section 3.3, certain adjustments to the parameter settings can have a significant impact on the condition number, while inducing relatively minor changes to the correlation function and hence to $\sigma_a^*$. Moreover, previous studies have shown that even an approximate correlation structure in $\mathbf{R}$ can yield higher quality analyses than ones obtained with wrongly assuming uncorrelated observation errors (e.g., Reference 12).

Corollaries 1,2 and Figure 6 can be used as a guideline to find suitable parameters. In this scenario, we would like to pick values of $M_o$ and $D_o$ that are 'close enough' to our target parameter values of $M_o = 10$ and $D_o = 120$ km so as not to increase the analysis error by too much at full convergence ($\sigma_a^*$), but which produce a much smaller condition number for $\mathbf{S}$. In Figure 9, we consider three additional experiments (labeled $\widetilde{\mathbf{R}}_3$, $\widetilde{\mathbf{R}}_4$, and $\widetilde{\mathbf{R}}_5$), which all use correlation models that are approximate compared to the actual one used to generate the observation error ($M_o = 10$ and $D_o = 120$ km) but which lead to improved convergence rates. Figure 9 shows that all three experiments outperform the diagonal $\mathbf{R}$ experiments $\widetilde{\mathbf{R}}_1$ and $\widetilde{\mathbf{R}}_2$ at every iteration. We now discuss the choice of the parameter values for these experiments in relation to Corollaries 1,2 and Figure 6.

In the situation where using the accurate correlation model would degrade the condition number, for given values of $M_o$, $M_b$, and $D_b$, we can modify $D_o$ to approximate the theoretical minimum condition number predicted by Corollaries 1 or 2. The experiment with $\widetilde{\mathbf{R}}_3$ corresponds to the 'extreme' case where $D_o$ is modified using Equation (29) ($M_o > M_b$ in this experiment) so its value coincides exactly with the minimum. To do so, we retain the true value of $M_o = 10$ but use an approximate value of $D_o = 50$ km to compute the analysis, instead of the length-scale of 120 km that was used to generate the correlated observation errors. The consequence of reducing $D_o$ and not $M_o$ is that the correlation function

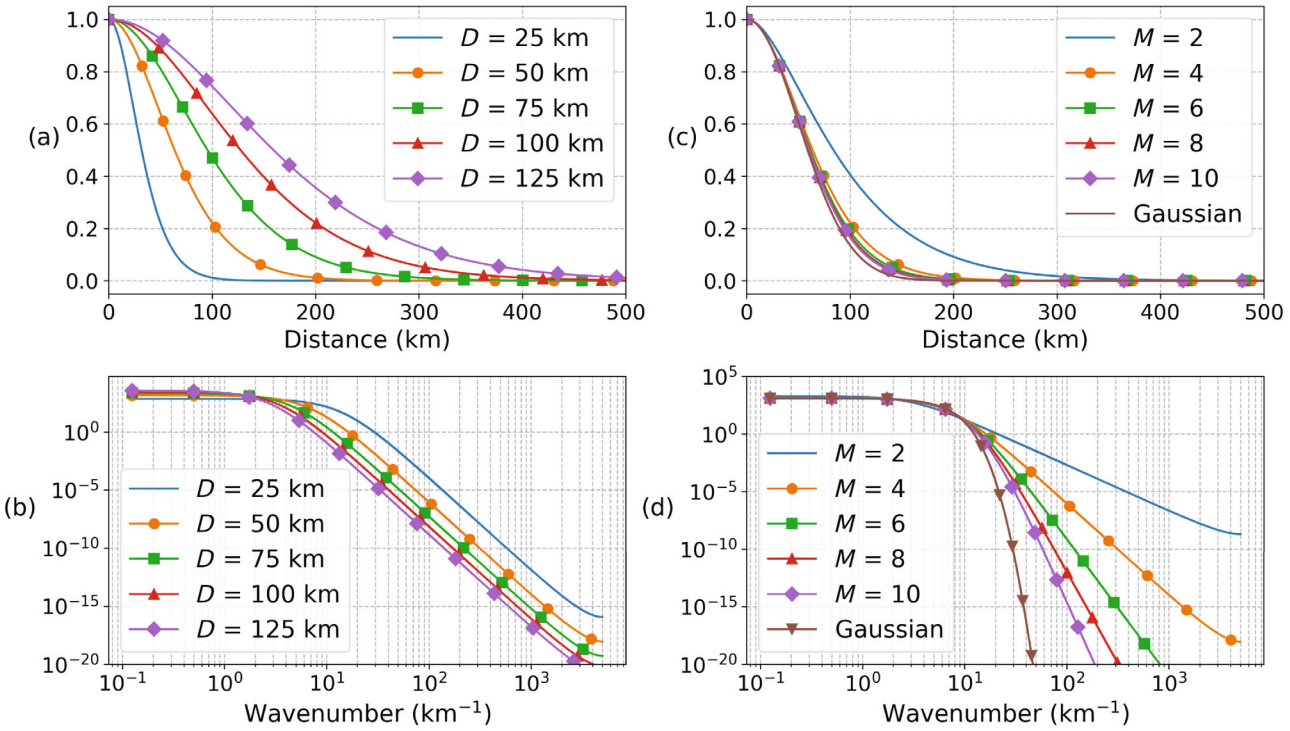



**FIGURE A1** (a) Equation (A2) plotted as a function of distance $r$, and (b) Equation (A4) plotted as a function of wavenumber. The curves are displayed for different values of $D$ and fixed value of $M = 4$. Panels (c) and (d) show corresponding plots where $M$ is varied for fixed value of $D = 50$ km.

associated with $\widetilde{\mathbf{R}}_3$ has the same shape as that of $\mathbf{R}$ but covers a much smaller range (see Figure A1a). Decreasing $D_o$ to this value reduces the condition number by a factor of $10^6$. With these new parameters, $\chi = 10^{-2}$, and the condition number obtained with $\widetilde{\mathbf{R}}_3$ is lower than the one obtained with either $\widetilde{\mathbf{R}}_1$ or $\widetilde{\mathbf{R}}_2$ (the condition number with $\widetilde{\mathbf{R}}_2$ is only 17 times lower than the condition number with $\widetilde{\mathbf{R}}_1$). Figure 9b shows that the experiment with $\widetilde{\mathbf{R}}_3$ outperforms both diagonal approximations at every iteration. With this modified value of $D_o$, the error reduction is 27% compared to 35% with the actual value of $D_o$, but allows a much faster convergence.

Another possibility is to modify $M_o$, so that a smaller modification on $D_o$ is required to approximate a minimum condition number predicted by Corollaries 1 or 2. When $M_b = 8$ and $D_b = 60$ km, Equation (29) associated with Corollary 1 predicts a minimum with the parameter pairs ($M_o = 10, D_o = 50$ km) and ($M_o = 8, D_o = 60$ km). It is thus possible to reach a minimum condition number with a smaller decrease of $D_o$ if $M_o$ is reduced from 10 to 8. The experiment with $\widetilde{\mathbf{R}}_4$ uses $M_o = M_b = 8$ and $D_b = D_o = 60$ km, which means that the correlation functions associated with $\mathbf{B}$ and $\mathbf{R}$ are very close (although $\mathbf{B}$ and $\mathbf{R}$ are not equal as there are fewer observations than grid points). Consequently, the condition number with $\widetilde{\mathbf{R}}_4$ is slightly lower than with $\widetilde{\mathbf{R}}_3$ (and thus also lower than with $\widetilde{\mathbf{R}}_1$ and $\widetilde{\mathbf{R}}_2$). This correlation model allows a slightly better error reduction than with $\widetilde{\mathbf{R}}_3$ (30% compared to 27%), while also converging slightly faster. As for the experiment with $\widetilde{\mathbf{R}}_3$, intermediate values of $D_o$ which approach the local minima while staying closer to the true parameters might offer a better compromise between a fast convergence and good error reduction at full convergence.

In the two previous experiments, the parameter values were chosen in order to reach a local condition number minimum predicted by Corollaries 1 or 2. If the estimated parameter values have a severe impact on convergence as in Figure 8 then Figure 6 suggests that, to lower the condition number, we should reduce the values of $M_o$ and/or $D_o$. In particular, if $M_o < M_b$ and $D_o < D_b$ then $\chi < 1$. Generally speaking, lower values of $D_o$ or $M_o$, relative to the corresponding values of $D_b$ and $M_b$, reduce the risk of the condition number being increased compared to the condition number with a diagonal $\mathbf{R}$ (i.e., of being in the red area of Figure 6 where $\chi > 1$). The parameters $D_o$ and $M_o$ can be reduced progressively through trial-and-error to determine a convergence rate at least as good as the one obtained with $\widetilde{\mathbf{R}}_1$. In particular, reducing $M_o$ does not necessarily induce a visible change in the shape of the correlation function (Figure A1b) but has a large effect on its Fourier transform at high wavenumbers (Figure A1d). This correlation function is associated with the first row of the correlation matrix, which is circulant. The Fourier transform of the correlation function thus corresponds

to the eigenvalues of the correlation matrix (see Section 3.1). The Fourier transform of the correlation function can be understood as the power spectrum associated with the observation-error correlation matrix (which does not necessarily correspond to the true observation-error correlation matrix used to simulate the observations). As a consequence, decreasing $M_o$ means that the observation-error power spectrum that we specify through the correlation matrix has larger variance at high wavenumbers than that of the actual observation-error power spectrum.

For example, in the experiment with $\widetilde{\mathbf{R}}_5$, we set $M_o = 2$ instead of 10 while keeping the correct value of $D_o = 120$ km. Although this does not match the true observation-error correlations, the modified observation-error spectrum does not induce a large increase of the analysis error: the experiment with $\widetilde{\mathbf{R}}_5$ achieves an error reduction of 33%, which is close to the theoretical minimum of 35% and better than what is obtained with either $\widetilde{\mathbf{R}}_3$ or $\widetilde{\mathbf{R}}_4$. The low sensitivity of the analysis error to this large change in the specification of the observation error can be explained by the fact that this change mostly affects the high wavenumbers where both the observation error and background error (and hence analysis error) are negligible. A misspecification of the observation error at these wavenumbers is thus unlikely to contribute significantly to the total analysis error. However, changing the observation-error spectrum at high wavenumbers by reducing $M_o$ has a drastic effect on the conditioning of the Hessian matrix: with $\widetilde{\mathbf{R}}_5$, the condition number is approximately 20 times smaller than with $\widetilde{\mathbf{R}}_1$ (and thus slightly smaller than with $\widetilde{\mathbf{R}}_2$). This large difference in the condition number can be explained by the fact that with $M_o$ equal to 10, the minimum eigenvalue of $\mathbf{R}$ is extremely small, which can significantly increase the amplitude of the maximum eigenvalue of the Hessian matrix according to Lemma 2. Reducing $M_o$ can increase the amplitude of the minimum eigenvalue by several orders of magnitude (see the end of the spectrum in Figure A1d, or Equation 17 with $i = m - 1$), and thus reduce the condition number without altering significantly the observation-error spectrum at low wavenumbers. If a case of poor convergence were to be observed in a real application, then this result suggests that reducing $M_o$ until an acceptable convergence rate is reached would be a reasonable remedy.

## 5 | SUMMARY AND CONCLUSIONS

Data assimilation concerns the problem of determining the optimal state of a system given observations, a background (prior) estimate of the state, and constraints that link the system state to the observations. Mathematically, the problem can be cast as one of nonlinear weighted least-squares. The technique of variational data assimilation, which is commonly used in atmospheric and ocean applications, seeks an approximate solution by using a truncated Gauss–Newton (TGN) algorithm to minimize the cost function of the nonlinear weighted least-squares problem. The TGN algorithm approximates the nonlinear problem by a connected sequence of linear sub-problems where each sub-problem is solved using a conjugate gradient (CG) algorithm.

In this article, we have studied the convergence properties of the CG algorithm with respect to parameter specifications in the background-error covariance matrix ($\mathbf{B}$) and observation-error covariance matrix ($\mathbf{R}$) whose respective inverse matrices are used to define the weights for the background and observations in the cost function. In line with common practice in variational data assimilation, we considered a CG algorithm that uses $\mathbf{B}$ as a preconditioner, which we referred to as the $\mathbf{B}$-preconditioned conjugate gradient ($\mathbf{B}$-PCG) algorithm. Our results have shown that the convergence rate of $\mathbf{B}$-PCG (and thus of the whole TGN minimization) is highly sensitive to parameters controlling the shape of typical covariance functions used to model $\mathbf{B}$ and $\mathbf{R}$. In particular, the number of CG iterations needed to reach a given tolerance can change by a few orders of magnitude depending on the relative parameter specifications between $\mathbf{B}$ and $\mathbf{R}$. This underlines the importance of including convergence impact as an additional constraint when adjusting covariance model parameters to fit covariances estimated from statistics.

We began by recalling the general upper bounds on the condition number derived by Reference 30. These upper bounds do not depend on the type of covariance matrices used, and thus can be overly pessimistic in specific cases. In order to derive more accurate bounds, we need to consider specific covariance matrices. In this article, we have focussed on covariance matrices that can be modeled as a matrix-vector product using a diffusion operator. Diffusion operators are commonly used for modeling spatial covariances in ocean data assimilation[24] and are closely related to other techniques for modeling spatial covariances in atmospheric data assimilation,[45] geostatistics,[43] inverse problems,[56] and uncertainty quantification.[57] They are well suited for problems that have large state and observation vectors, they provide convenient access to an inverse covariance operator (specifically $\mathbf{R}^{-1}$ as required by $\mathbf{B}$-PCG), and they have useful flexibility for specifying covariance functions with different characteristics.

In order to simplify the theoretical analysis, we assumed that the basic parameters of the diffusion-based covariance models for both $\mathbf{B}$ and $\mathbf{R}$ were constant. These parameters consist of the standard deviations ($\sigma_b$ and $\sigma_o$), as well as

parameters that control the degree of smoothness (integers $M_b$ and $M_o$) and spatial range (length-scales $L_b$ and $L_o$) of the underlying correlation functions. (The quantities with subscripts 'b' and 'o' refer to the background and observation quantities, respectively.) With these assumptions, the covariance functions implied by the diffusion model are of Matérn type. In $\mathbb{R}$, they are auto-regressive (AR) functions of order $M$, which are the functions relevant for our one-dimensional (1D) analysis. Furthermore, we assumed that the grids supporting the background state and observations have uniform resolution $h_b$ and $h_o = \zeta h_b$ where $\zeta$ is a positive integer, which implies that there are fewer observations than background grid points. Under these assumptions, we derived an analytical expression for the eigenvalues of the **B**-preconditioned Hessian matrix (**S**). By further assuming that $\mathbf{HBH}^T$ can be approximated by a diffusion operator $\mathbf{B}_o$ that is discretized directly on the observation grid, it has been possible to derive criteria that the parameter pairs $(M_b, L_b)$ and $(M_o, L_o)$ must jointly satisfy to obtain a minimum upper bound for the condition number of **S**. These constraints are exact when the observation and background grids coincide ($\zeta = 1$), but are affected by a minor discretization error when the observation grid is coarser than the background grid ($\zeta < 1$). We used these analytical results to interpret numerical results from experiments with a 1D variational data assimilation system (1D-Var).

First, our analytical expressions expose the already well-known result that increasing (decreasing) the ratio $\sigma_b/\sigma_o$ leads to an increase (decrease) of the condition number of **S**. Furthermore, when $M_o = M_b$, our results show that the condition number is minimized when the same correlation model is used for both background and observation errors (i.e., $L_b = L_o$). This is consistent with the results of Tabeart et al.[30] who considered only the special case when $M_o = M_b = 2$ (i.e., when the correlation functions are second-order AR functions). However, the main contribution of our work has been to extend the analysis to the more realistic case where the background and observation errors are modeled with different smoothness parameters ($M_o \neq M_b$).

While we have derived the analytical results in terms of the parameter pairs $(M_b, L_b)$ and $(M_o, L_o)$, we have mainly interpreted them and the results of the numerical experiments in terms of the parameter pairs $(M_b, D_b)$ and $(M_o, D_o)$ where $D_b$ and $D_o$ are alternative ('Daley') length-scale parameters commonly used for differentiable correlation functions in data assimilation.[39] Specifically, for the 1D problem under consideration, $D_b = L_b\sqrt{2M_b - 3}$ and $D_o = L_o\sqrt{2M_o - 3}$ where $M_b > 1$ and $M_o > 1$. In terms of fixed values of $D_b$ and $D_o$, the AR functions have the convenient property that they converge to Gaussian functions for large $M_b$ and $M_o$. Our results have also exposed a direct relationship with closely-related ('Stein') length-scale parameters, $\rho_b = L_b\sqrt{2M_b - 1}$ and $\rho_o = L_o\sqrt{2M_o - 1}$, used in geostatistics.[41] In terms of fixed values of $\rho_b$ and $\rho_o$, the AR functions also converge to Gaussian functions for large $M_b$ and $M_o$, and are defined for both the differentiable ($M_b > 1$ and $M_o > 1$) and non-differentiable AR functions ($M_b = M_o = 1$).

The condition number is markedly more sensitive to the parameter specifications for the case $M_o > M_b$ than $M_o < M_b$. This has been illustrated in the numerical experiments and is evident from the analytical expression (Equation 29) that describes the relationship between $(M_b, D_b)$ and $(M_o, D_o)$ required to achieve the minimum upper bound of the condition number when $M_o > M_b$. In general, $D_o$ needs to be much smaller than $D_b$ for this optimality condition to be met because of the presence of $M_b$ and $M_o$ as exponents in the expressions. For this case, the small eigenvalues of **S** are amplified by $\mathbf{R}^{-1}$ more than they are damped by **B**, which can result in a drastic increase of the condition number and thus a significant risk that the convergence rate of **B**-PCG will be substantially degraded. For the case $M_o < M_b$, the minimum upper bound of the condition number is attained when the 'Stein' length-scales $\rho_b$ and $\rho_o$ are equal (Equation 26). In contrast with the case $M_o > M_b$, this means that the minimum upper bound is obtained when the background- and observation-error correlation functions have similar spatial range. When $M_o < M_b$ and $D_o \leq D_b$, accounting for observation-error correlations systematically improves the conditioning of **S** compared to the case when a diagonal **R** is used.

While $M_o$ and $M_b$ are intended as free parameters of the correlation model, to be adjusted to achieve the best possible fit to statistical estimates of correlated error, they also provide valuable leverage for controlling the conditioning of **S** when a non-diagonal **R** is used. Whereas using a non-diagonal **R** is likely to degrade significantly the convergence rate when $M_o > M_b$, it can accelerate the convergence rate compared to the case where a diagonal **R** is used when $M_o < M_b$. In practice, this situation would correspond to choosing a Gaussian-like correlation function for background error (e.g., $M_b \approx 10$) and a correlation function with fatter tails (more power at smaller scales) for observation error (e.g., $M_o = 2$). Interestingly, there is evidence in the atmospheric data assimilation literature that suggests that error correlations for certain observation types do exhibit fat tails. In this case, the interest in using a non-diagonal **R** is twofold: it can accelerate the **B**-PCG convergence rate as well as providing a more accurate correlation model. For the case where statistical estimates of the correlation parameters result in unfavourable values in terms of conditioning ($M_o > M_b$ and/or $D_o \gg D_b$), **S** can be 'reconditioned' by adjusting the values of $M_o$ and $D_o$ to enable faster convergence. We have shown in our 1D-Var experiments that approximate correlation models can be used to reduce the condition number without causing a significant loss of solution accuracy at full convergence. This corroborates the conclusion of several previous studies (e.g.,

Reference 12) that even a very crude approximation of the observation-error correlations can give a better solution than one obtained by ignoring them altogether.

The analysis in this article has exposed important sensitivities of the convergence rate of **B**-PCG to fundamental parameters of diffusion model representations of **B** and **R**, and has lead to conditions for adjusting the parameters to improve the conditioning of **S**. We can expect similar results in higher dimensions where diffusion kernels have similar (Matérn-like) functional forms as those in our 1D study. However, more work is required to extend these results to account for more sophisticated diffusion models, such as ones that include diffusion tensors that are anisotropic and spatially varying, or multiple-scale and hybrid formulations that are built from linear combinations of diffusion operators. In this study, we considered a simple observation network. The convergence properties need to be revisited in an operational-like framework using a full network of diverse observations for which only a subset may be affected by spatially correlated errors in **R**, and where observation operators will be much more complex. The results from this study are a first step towards understanding and controlling the convergence properties in this more challenging framework.

## ACKNOWLEDGMENTS

This work has benefitted from funding from the Copernicus Climate Change Service (C3S), one of six services of the European Union's Copernicus Programme, which is implemented by ECMWF on behalf of the European Commission. It is a contribution to the contract C3S_601_INRIA: "*Advancing ocean data assimilation methodology for climate applications*". Additional support has been provided by the French National Programme LEFE/INSU.

## CONFLICT OF INTEREST STATEMENT

This study does not have any conflicts to disclose.

## DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

## ENDNOTES

[1]In the current context, the time coordinate in the diffusion equation does not represent physical time but should be interpreted as a pseudo-time coordinate that controls the smoothing properties of the diffusion kernel. This explains why 'time' has been written within quotation marks.

[2]$w = \sqrt{2}$ in Reference 41 [pp. 48–50], $w = 2$ in Reference 43, and $w = 1$ in Reference 42 [Chap. 4.2].

[3]The exact condition number has been computed using results from Theorem 3 described later in Section 3.

[4]Each column (row) of a circulant matrix is a cyclic permutation of the previous column (row).

## ORCID

*Olivier Goux* https://orcid.org/0000-0002-2848-7442
*Youssef Diouane* https://orcid.org/0000-0002-6609-7330

## REFERENCES

1. Gratton S, Lawless AS, Nichols NK. Approximate Gauss–Newton methods for nonlinear least squares problems. SIAM J Optim. 2007;18:106–32.
2. Courtier P, Thépaut JN, Hollingsworth A. A strategy for operational implementation of 4D-Var, using an incremental approach. Q J Roy Meteorol Soc. 1994;120:1367–87.
3. Gürol S, Weaver AT, Moore AM, Piacentini A, Arango HG, Gratton S. **B**-preconditioned minimization algorithms for variational data assimilation with the dual formulation. Q J Roy Meteorol Soc. 2013;140:539–56.
4. Derber J, Rosati A. A global oceanic data assimilation system. J Phys Oceanogr. 1989;19:1333–47.
5. Lorenc AC. Development of an operational variational assimilation scheme. J Meteorol Soc Japan. 1997;75:339–46.
6. Bormann N, Bauer P. Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I: methods and application to ATOVS data. Q J Roy Meteorol Soc. 2010;136:1036–50.
7. Waller JA, Ballard S, Dance S, Kelly G, Nichols NK, Simonin D. Diagnosing horizontal and inter-channel observation error correlations for SEVIRI observations using observation-minus-background and observation-minus-analysis statistics. Remote Sens (Basel). 2016;8:581.
8. Liu ZQ, Rabier F. The interaction between model resolution, observation resolution and observation density in data assimilation: a one-dimensional study. Q J Roy Meteorol Soc. 2002;128:1367–86.
9. Reid R, Good S, Martin MJ. Use of uncertainty inflation in OSTIA to account for correlated errors in satellite-retrieved sea surface temperature data. Remote Sens (Basel). 2020;12:1083.

10. Rainwater S, Bishop CH, Campbell WF. The benefits of correlated observation errors for small scales. Q J Roy Meteorol Soc. 2015;141:3439–45.
11. Healy SB, White AA. Use of discrete Fourier transforms in the 1D-Var retrieval problem. Q J Roy Meteorol Soc. 2005;131:63–72.
12. Stewart LM, Dance SL, Nichols NK. Data assimilation with correlated observation errors: experiments with a 1-D shallow water model. Tellus Ser A. 2013;65:19546.
13. Ruggiero GA, Cosme E, Brankart JM, Le Sommer J. An efficient way to account for observation error correlations in the assimilation of data from the future SWOT high-resolution altimeter mission. J Atmos Oceanic Tech. 2016;33:2755–68.
14. Pinnington EM, Casella E, Dance SL, et al. Investigating the role of prior and observation error correlations in improving a model forecast of forest carbon balance using four-dimensional variational data assimilation. Agric Meteorol. 2016;228-229:299–314.
15. Brankart JM, Ubelmann C, Testut CE, Cosme E, Brasseur P, Verron J. Efficient parameterization of the observation error covariance matrix for square root or ensemble Kalman filters: application to ocean altimetry. Mon Weather Rev. 2009;137:1908–27.
16. Michel Y. Revisiting Fisher's approach to the handling of horizontal spatial correlations of observation errors in a variational framework. Q J Roy Meteorol Soc. 2018;144:2011–25.
17. Bédard J, Buehner M. A practical assimilation approach to extract smaller-scale information from observations with spatially correlated errors: an idealized study. Q J Roy Meteorol Soc. 2019;146:468–82.
18. Guillet O, Weaver AT, Vasseur X, Michel Y, Gratton S, Gürol S. Modelling spatially correlated observation errors in variational data assimilation using a diffusion operator on an unstructured mesh. Q J Roy Meteorol Soc. 2019;145:1947–67.
19. Hu G, Dance SL. Efficient computation of matrix-vector products with full observation weighting matrices in data assimilation. Q J Roy Meteorol Soc. 2021;147:4101–21.
20. Guttorp P, Gneiting T. Studies in the history of probability and statistics XLIX on the Matérn correlation family. Biometrika. 2006;93:989–95.
21. Mirouze I, Weaver AT. Representation of correlation functions in variational assimilation using an implicit diffusion operator. Q J Roy Meteorol Soc. 2010;136:1421–43.
22. Weaver AT, Mirouze I. On the diffusion equation and its application to isotropic and anisotropic correlation modelling in variational assimilation. Q J Roy Meteorol Soc. 2013;139:242–60.
23. Egbert G, Bennett A, Foreman M. Topex/Poseidon tides estimated using a global inverse model. J Geophys Res. 1994;99:24821–52.
24. Weaver AT, Courtier P. Correlation modelling on the sphere using a generalized diffusion equation. Q J Roy Meteorol Soc. 2001;127:1815–46.
25. Weaver AT, Tshimanga J, Piacentini A. Correlation operators based on an implicitly formulated diffusion equation solved with the Chebyshev iteration. Q J Roy Meteorol Soc. 2015;142:455–71.
26. Weaver AT, Chrust M, Ménétrier B, Piacentini A. An evaluation of methods for normalizing diffusion-based covariance operators in variational data assimilation. Q J Roy Meteorol Soc. 2020;147:289–320.
27. Axelsson O, Kaporin I. On the sublinear and superlinear rate of convergence of conjugate gradient methods. Numer Algorithms. 2000;25:1–22.
28. Haben SA, Lawless AS, Nichols NK. Conditioning of incremental variational data assimilation, with application to the met office system. Tellus Ser A. 2011;63:782–92.
29. Tabeart JM, Dance SL, Haben SA, Lawless AS, Nichols NK, Waller JA. The conditioning of least-squares problems in variational data assimilation. Numer Linear Algebra Appl. 2018;25:e2165.
30. Tabeart JM, Dance SL, Lawless AS, Nichols NK, Waller JA. New bounds on the condition number of the hessian of the preconditioned variational data assimilation problem. Numer. Linear Algebra Appl. 2021;29:e2405.
31. Ide K, Courtier P, Ghil M, Lorenc AC. Unified notation for data assimilation: operational, sequential and variational. J. Meteorol. Soc. Japan. 1997;75B:181–9.
32. Golub GH, Van Loan CF. Matrix computations. Baltymore, MD: J. Hopkins Uni. Press; 2013.
33. Saad Y. Iterative methods for sparse linear systems. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2003.
34. Nocedal J, Wright S. Numerical optimization. New York: Springer-Verlag GmbH; 2006.
35. Kelley CT. Iterative methods for linear and nonlinear equations. Cambridge, UK: Cambridge University Press; 1987.
36. Axelsson O. Iterative solution methods. Cambridge, UK: Cambridge University Press; 1994.
37. Lorenc AC. Optimal nonlinear objective analysis. Q J Roy Meteorol Soc. 1988;114:205–40.
38. Whittle P. Stochastic processes in several dimensions. Bull Inst Int Stat. 1963;40:974–94.
39. Daley R. Atmospheric data analysis. Cambridge, UK: Cambridge University Press; 1991.
40. Pannekoucke O, Berre L, Desroziers G. Background-error correlation length-scale estimates and their sampling statistics. Q J Roy Meteorol Soc. 2008;134:487–508.
41. Stein LM. Interpolation of spatial data. New York, NY: Springer-Verlag; 1999.
42. Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Cambridge, MA: MIT Press; 2006.
43. Lindgren F, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J Roy Stat Soc: Series B Stat Method. 2011;73:423–98.
44. Gaspari G, Cohn SE. Construction of correlation functions in two and three dimensions. Q J Roy Meteorol Soc. 1999;125:723–57.
45. Purser RJ, Wu WS, Parrish DF, Roberts NM. Numerical aspects of the application of recursive filters to variational statistical analysis. Part I: spatially homogeneous and isotropic Gaussian covariances. Mon Weather Rev. 2003;131:1524–35.

46. Derber J, Bouttier F. A reformulation of the background error covariance in the ECMWF global data assimilation system. Tellus A: Dyn Meteorol Oceanography. 1999;51:195–221.
47. Weaver AT, Deltel C, Machu E, Ricci S, Daget N. A multivariate balance operator for variational ocean data assimilation. Q J Roy Meteorol Soc. 2005;131:3605–25.
48. Waller JA, Simonin D, Dance SL, Nichols NK, Ballard SP. Diagnosing observation error correlations for Doppler radar radial winds in the met Office UKV model using observation-minus-background and observation-minus-analysis statistics. Mon Weather Rev. 2016;444:3534–51.
49. Tarantola A. Inverse problem theory: methods for data fitting and model parameter estimation. Amsterdam: Elsevier; 1987.
50. Gray RM. Toeplitz and circulant matrices: a review. Found Trends Commun Inf Theory. 2005;2:155–239.
51. Harville DA. Matrix algebra from a statistician's perspective. New York: Springer; 1997.
52. Andersson E, Fisher M, Munro R, McNally A. Diagnosis of background errors for radiances and other observable quantities in a variational data assimilation scheme, and the explanation of a case of poor convergence. Q J Roy Meteorol Soc. 2000;126:1455–72.
53. Weston PP, Bell W, Eyre JR. Accounting for correlated error in the assimilation of high-resolution sounder data. Q J Roy Meteorol Soc. 2014;140:2420–9.
54. Tabeart JM, Dance SL, Lawless AS, et al. The impact of using reconditioned correlated observation-error covariance matrices in the met office 1D-Var system. Q J Roy Meteorol Soc. 2020;146:1372–90.
55. Gelb A. Applied optimal estimation. Cambridge: The MIT Press; 1974.
56. Bui-Thanh T, Ghattas O, Martin J, Stadler G. A computational framework for infinite-dimensional Bayesian inverse problems. Part I: the linearized case, with application to global seismic inversion. SIAM J. Sci. Comput. 2013;35:A2494–523.
57. Gmeiner B, Drzisga D, Rüde U, Scheichl R, Wohlmuth B. Scheduling massively parallel multigrid for multilevel Monte Carlo methods. SIAM J Sci Comput. 2017;39:S873–97.

**How to cite this article:** Goux O, Gürol S, Weaver AT, Diouane Y, Guillet O. Impact of correlated observation errors on the conditioning of variational data assimilation problems. Numer Linear Algebra Appl. 2024;31(1):e2529. https://doi.org/10.1002/nla.2529

## APPENDIX A. MATéRN FUNCTIONS AND DIFFUSION OPERATORS ON $\mathbb{R}$ AND $\mathbb{S}$

### A.1 Diffusion on $\mathbb{R}$

Matérn random fields on $\mathbb{R}^d$ can be derived by solving a general stochastic fractional partial differential equation (PDE).[20,38] Here, we are interested in the correlation functions of a subset of Matérn fields on $\mathbb{R}$ ($d = 1$) where parameters are chosen such that the generating PDE has a simplified form for numerical computations.

Let $\chi : z \mapsto \chi(z)$ and $\eta : z \mapsto \eta(z)$ be square-integrable functions ($\chi, \eta \in L^2(\mathbb{R})$) of the spatial coordinate $z \in \mathbb{R}$. We consider solutions of the following elliptic equation on $\mathbb{R}$:

$$\frac{1}{\gamma^2}\left(I - L^2\frac{\partial^2}{\partial z^2}\right)^M \eta(z) = \chi(z), \tag{A1}$$

where $I$ is the identity operator, $M$ is a positive integer, $L$ is a length-scale parameter, and $\gamma^2$ is a normalization constant. Equation (A1) can be interpreted as the *inverse* of a diffusion operator, $\eta \mapsto \mathcal{L}^{-1}\eta$, which is formed by discretizing the time derivative of the diffusion equation with an Euler backward (implicit) scheme and by applying the resulting operator over $M$ time steps.[21] With this interpretation, $L^2 = \mu\Delta t$ where $\mu$ is the diffusion coefficient and $\Delta t$ is the time step. The integral solution of Equation (A1) is thus a diffusion operator, $\chi \mapsto \mathcal{L}\chi$. The solution, which is straightforward to derive using the Fourier transform, is a convolution operator, $\mathcal{L}\chi \equiv c \times \chi$, where $c = c(r)$ is an $M$th-order AR function (a polynomial times the exponential function) given by

$$c(r) = \sum_{j=0}^{M-1} \beta_j \left(\frac{r}{L}\right)^j e^{-r/L}, \tag{A2}$$

$r = |z - z'|$ is the Euclidean distance between points $z$ and $z'$, and

$$\beta_j = \frac{2^j (M-1)!\,(2M-j-2)!}{j!\,(M-j-1)!\,(2M-2)!}.$$

Setting the normalization constant to

$$\gamma^2 = \nu\, L,$$

where

$$\nu = \frac{2^{2M-1}[(M-1)!]^2}{(2M-2)!}, \tag{A3}$$

ensures that $c(0) = 1$.[21] The power spectrum of $c$, which is given by the Fourier transform $\hat{c}$ of $c$, describes the smoothness properties of $c$ as a function of spectral scale:

$$\hat{c}(\hat{z}) = \frac{\gamma^2}{\left(1 + L^2\hat{z}^2\right)^M}, \tag{A4}$$

where $\hat{z}$ is the spectral wavenumber.

We focus on the differentiable AR functions that correspond to $M > 1$. For these functions, we use a standard parameter[39]

$$D = \sqrt{-\frac{1}{\partial^2 c/\partial z^2|_{z=z'}}},$$

to characterize the length-scale of the correlation function. The parameter $D$, which we call the Daley length-scale, corresponds to the distance between $z = z'$ and the mid-amplitude point of a parabola that osculates the AR function at $z = z'$. Using Equation (A2), it is straightforward to show that $D = L\sqrt{2M-3}$ (Equation 7), which is a function of both $L$ and $M$. An important property of AR functions is that, for fixed $D$, they converge to the Gaussian function $c_g(r)$ as $M \to \infty$:

$$c_g(r) = \exp\left(-r^2/2D^2\right).$$

Figure A1 shows the effect on $c$ and $\hat{c}$ of varying $D$ for a fixed value of $M$, and vice versa. Increasing $D$ with $M$ held fixed increases the spatial reach of the correlation functions but does not affect their spectral decay rate at small wavelengths. On the other hand, increasing $M$ with $D$ held fixed results in correlation functions with thinner tails and sharper spectral decay rates at small wavelengths.

### A.2 Diffusion on $\mathbb{S}$

References 29 and 30 use a SOAR function, which is equal to Equation (A2) with $M = 2$ and hence $D = L$ from Equation (7). Furthermore, they restrict the SOAR function to the circular domain ($\mathbb{S}$) of radius $a$ by using chordal distance $r = 2a\sin(\theta/2)$ where $\theta$ is the angle between points $z$ and $z'$ on the circle. This ensures that $c(r)$ is positive definite on $\mathbb{S}$.[44] Taking $a$ as the radius of the Earth, the domain $\mathbb{S}$ can be interpreted as a latitude circle at the Equator.

In this article, we have also considered a circular domain of radius $a$. For length-scales $L \ll a$, the correlation functions associated with the diffusion operator applied on $\mathbb{S}$ are approximately Matérn since the influence of curvature is minor. It is instructive nevertheless to present the exact correlation functions on $\mathbb{S}$, which can be derived by considering the solution of the elliptic equation

$$\frac{1}{\gamma^2}\left(I - \frac{L^2}{a^2}\frac{\partial^2}{\partial\phi^2}\right)^M \eta(\phi) = \mu(\phi), \tag{A5}$$

subject to periodic boundary conditions on the solution and its derivative:

$$\eta(-\pi) = \eta(\pi),$$
$$\frac{\partial\eta}{\partial\phi}\Big|_{\phi=-\pi} = \frac{\partial\eta}{\partial\phi}\Big|_{\phi=\pi}.$$

Solving Equation (A5) is equivalent to solving Equation (A1) on the periodic domain $-Z \le z \le Z$ with $z = a\phi$ and $Z = a\pi$.

The solutions that satisfy the boundary conditions are of the general form

$$\eta(\phi) = \sum_{m=0}^{\infty} A_m \cos(m\phi) + B_m \sin(m\phi). \tag{A6}$$

The coefficients $A_m$ and $B_m$ can be determined using the orthogonality relations of the sine and cosine functions:

$$\begin{aligned}\int_{-\pi}^{\pi} \sin(m\phi)\sin(n\phi)\mathrm{d}\phi &= \int_{-\pi}^{\pi} \cos(m\phi)\cos(n\phi)\mathrm{d}\phi = \pi\ \delta_{mn},\\ \int_{-\pi}^{\pi} \sin(m\phi)\cos(n\phi)\mathrm{d}\phi &= 0\ \ \forall\ m, n,\end{aligned}$$

where $\delta_{mn}$ is the Kronecker delta. To determine $A_m$, we substitute Equations (A6) in (A5), multiply the resulting equation by $\cos(n\phi)$, integrate from $-\pi$ to $\pi$, and use the orthogonality relations above. This yields

$$A_m = \frac{\gamma^2}{\pi}\left(1 + \frac{L^2}{a^2}\ m^2\right)^{-M} \int_{-\pi}^{\pi} \mu(\phi)\cos(m\phi)\ \mathrm{d}\phi. \tag{A7}$$

To determine $B_m$, we follow the same procedure but multiply by $\sin(n\phi)$. This yields

$$B_m = \frac{\gamma^2}{\pi}\left(1 + \frac{L^2}{a^2}\ m^2\right)^{-M} \int_{-\pi}^{\pi} \mu(\phi)\sin(m\phi)\ \mathrm{d}\phi. \tag{A8}$$

Substituting Equations (A7) and (A8) into (A6), and using the trigonometric identity

$$\cos(m(\phi - \phi')) = \cos(m\phi)\ \cos(m\phi') + \sin(m\phi)\ \sin(m\phi'),$$

yields the solution

$$\eta(\phi) = \int_{-\pi}^{\pi} c(\theta)\ \mu(\phi')\ \mathrm{d}\phi',$$

where $\theta = \phi - \phi'$,

$$c(\theta) = \gamma^2 \sum_{m=0}^{\infty} c_m \cos(m\theta), \tag{A9}$$

and

$$c_m = \frac{1}{\pi}\left(1 + \frac{L^2}{a^2}\ m^2\right)^{-M}. \tag{A10}$$

The normalization factor and Daley length-scale are, respectively,

$$\gamma^2 = \frac{1}{\sum_{m=0}^{\infty} c_m},$$

and

$$D = a\sqrt{-\frac{1}{\partial^2 c/\partial\phi^2|_{\theta=0}}} = a\ \sqrt{\frac{1}{\sum_{m=0}^{\infty} m^2 c_m}}.$$

All valid continuous isotropic correlation functions on $\mathbb{S}$ can be represented by a Fourier cosine series expansion with non-negative Fourier coefficients (see Theorem 2.11 in Reference 44), which is clearly satisfied by Equations (A9)

and (A10). The smoothness properties of the correlation function are determined by the Fourier coefficients $c_m$ in Equation (A10). They can be seen to have a similar dependence on $L$ and $M$ as $\hat{c}$ in Equation (A4) where we can associate $\hat{z}$ on $\mathbb{R}$ with $m^2/a^2$ on $\mathbb{S}$.

## APPENDIX B. PROOF OF THEOREM 5

The eigenvalues of $\mathbf{S}_o$ are bounded below by 1 (see Theorem 4), which implies that

$$\kappa(\mathbf{S}_o) \leq \max_{i \in [\![0, m-1]\!]} \lambda_i(\mathbf{S}_o) = \lambda_{\max}(\mathbf{S}_o).$$

Since $0 \leq \sin^2(y) \leq 1$ for any $y \in [0, \pi]$, $\lambda_{\max}(\mathbf{S}_o)$ is bounded by

$$\lambda_{\max}(\mathbf{S}_o) \leq \max_{x \in [0,1]} \phi(x),$$

where $\phi$ is a continuously differentiable function given by

$$\phi(x) = 1 + \alpha \frac{\left[1 + 4\widetilde{L}_o^2 x\right]^{M_o}}{\left[1 + 4\widetilde{L}_{b/o}^2 x\right]^{M_b}}. \tag{B1}$$

We seek a solution to the following bound-constraint problem:

$$\max_{x \in [0,1]} \phi(x). \tag{B2}$$

Let $x^* \in [0, 1]$ be a stationary point for problem (B2) and let us first assume that such point is inside the domain; that is, $\phi'(x^*) = 0$. The derivative of the function $\phi$ can be expressed as

$$\phi'(x) = 4\alpha \; v(x) \; w(x),$$

where

$$v(x) = \left( \frac{\widetilde{L}_o^2 M_o}{1 + 4\widetilde{L}_o^2 x} - \frac{\widetilde{L}_{b/o}^2 M_b}{1 + 4\widetilde{L}_{b/o}^2 x} \right) \quad \text{and} \quad w(x) = \frac{\left[1 + 4\widetilde{L}_o^2 x\right]^{M_o}}{\left[1 + 4\widetilde{L}_{b/o}^2 x\right]^{M_b}}.$$

Since $w(x)$ is strictly positive, and $\alpha > 0$, a stationary point inside the domain satisfies $v(x^*) = 0$. This yields

$$x^* = \frac{\widetilde{L}_{b/o}^2 M_b - \widetilde{L}_o^2 M_o}{4\widetilde{L}_o^2 \widetilde{L}_{b/o}^2 (M_o - M_b)}. \tag{B3}$$

The second derivative of $f$ is

$$\phi''(x) = 16\alpha \; w(x) \left( \frac{\widetilde{L}_{b/o}^4 M_b}{\left(1 + 4\widetilde{L}_{b/o}^2 x\right)^2} - \frac{\widetilde{L}_o^4 M_o}{\left(1 + 4\widetilde{L}_o^2 x\right)^2} \right) + 16\alpha \; v(x)^2 \; w(x). \tag{B4}$$

Substituting (B3) into (B4) gives

$$\phi''(x^*) = 16\alpha (M_o - M_b) \; w(x^*) \left( \frac{\widetilde{L}_{b/o}^4 \widetilde{L}_o^4 (M_o - M_b)^2}{\left(\widetilde{L}_{b/o}^2 - \widetilde{L}_o^2\right)^2 M_o M_b} \right).$$

Therefore, the stationary point $x^*$ can be a maximum point if and only if $\phi''(x^*) < 0$; that is, if $M_o < M_b$. In addition, for $x^*$ to be a feasible point then $0 < x^* < 1$ and from Equation (B3) the following conditions must be satisfied:

$$\widetilde{L}_{b/o}^2 M_b - \widetilde{L}_o^2 M_o < 0,$$

and

$$\widetilde{L}_{b/o}^2 M_b - \widetilde{L}_o^2 M_o > 4\widetilde{L}_o^2 \widetilde{L}_{b/o}^2 (M_o - M_b).$$

For the other cases, $x^*$ is equal to either the lower bound ($x^* = 0$) or the upper bound ($x^* = 1$), with function values of

$$\phi(0) = 1 + \alpha,$$
$$\phi(1) = 1 + \alpha \frac{\left[1 + 4\widetilde{L}_o^2\right]^{M_o}}{\left[1 + 4\widetilde{L}_{b/o}^2\right]^{M_b}}.$$

Finally, substituting (B3) into (B1), we obtain that

$$\phi(x^*) = 1 + \alpha \left(\frac{\widetilde{L}_o^2}{M_b}\right)^{M_b} \left(\frac{M_o}{\widetilde{L}_{b/o}^2}\right)^{M_o} \left(\frac{M_b - M_o}{\widetilde{L}_o^2 - \widetilde{L}_{b/o}^2}\right)^{M_b - M_o}.$$

## APPENDIX C. PROOF OF COROLLARY 1

We consider $\eta$ as a function of $\widetilde{L}_o$, denoted $f(\widetilde{L}_o)$. Hereafter, the conditions (i), (ii), and (iii) will refer to the conditions stated in Theorem 5. We consider the case where condition (ii) does not hold, that is, $M_o \geq M_b$. In this case, Theorem 5 states that

$$f(\widetilde{L}_o) = 1 + \frac{\sigma_b^2 \nu_b \widetilde{L}_{b/o}}{\sigma_o^2 \nu_o \widetilde{L}_o} \max\left\{ \frac{\left(1 + 4\widetilde{L}_o^2\right)^{M_o}}{\left(1 + 4\widetilde{L}_{b/o}^2\right)^{M_b}};\ 1 \right\}. \tag{C1}$$

Let us first assume that $\left(1 + 4\widetilde{L}_o^2\right)^{M_o} < \left(1 + 4\widetilde{L}_{b/o}^2\right)^{M_b}$. Then, Equation (C1) simplifies to

$$f(\widetilde{L}_o) = 1 + \frac{\sigma_b^2 \nu_b \widetilde{L}_{b/o}}{\sigma_o^2 \nu_o \widetilde{L}_o},$$

which is a decreasing function of $\widetilde{L}_o$. Let us now assume that $\left(1 + 4\widetilde{L}_o^2\right)^{M_o} > \left(1 + 4\widetilde{L}_{b/o}^2\right)^{M_b}$. In this case, Equation (C1) becomes

$$f(\widetilde{L}_o) = 1 + \frac{\sigma_b^2 \nu_b \widetilde{L}_{b/o}}{\sigma_o^2 \nu_o \widetilde{L}_o} \frac{\left(1 + 4\widetilde{L}_o^2\right)^{M_o}}{\left(1 + 4\widetilde{L}_{b/o}^2\right)^{M_b}},$$

whose derivative is given by

$$f'(\widetilde{L}_o) = \underbrace{\frac{\sigma_b^2 \nu_b \widetilde{L}_{b/o}}{\sigma_o^2 \nu_o \widetilde{L}_o^2} \frac{\left(1 + 4\widetilde{L}_o^2\right)^{M_o - 1}}{\left(1 + 4\widetilde{L}_{b/o}^2\right)^{M_b}}}_{>0} \left(4\widetilde{L}_o^2 (2M_o - 1) - 1\right).$$

Under the assumption that $\widetilde{L}_o > 1/\left(2\sqrt{2M_o - 1}\right)$, $f(\widetilde{L}_o)$ is an increasing function of $\widetilde{L}_o$. As we already showed that $f(\widetilde{L}_o)$ is decreasing when $\left(1 + 4\widetilde{L}_o^2\right)^{M_o} < \left(1 + 4\widetilde{L}_{b/o}^2\right)^{M_b}$, the function $f$ then reaches its unique minimum when

$$\left(1 + 4\widetilde{L}_o^2\right)^{M_o} = \left(1 + 4\widetilde{L}_{b/o}^2\right)^{M_b}.$$

## APPENDIX D. PROOF OF COROLLARY 2

We consider $\eta$ as a function of $\widetilde{L}_o$, denoted by $f(\widetilde{L}_o)$. Hereafter, the conditions (i), (ii), and (iii) will refer to the conditions stated in Theorem 5. Let us first assume that condition (ii) holds, that is, $M_o < M_b$ whereas condition (i) does not hold, that is,

$$\widetilde{L}_o^2 M_o \leq \widetilde{L}_{b/o} M_b. \tag{D1}$$

From (D1), we first consider the variations of $f(\widetilde{L}_o)$ when $\widetilde{L}_o$ is in the interval $\left[0, \widetilde{L}_{b/o}\sqrt{M_b/M_o}\right]$. In this case, Theorem 5 states that

$$f(\widetilde{L}_o) = 1 + \frac{\sigma_b^2 \nu_b \widetilde{L}_{b/o}}{\sigma_o^2 \nu_o \widetilde{L}_o} \max\left\{\frac{(1 + 4\widetilde{L}_o^2)^{M_o}}{(1 + 4\widetilde{L}_{b/o}^2)^{M_b}};\ 1\right\}. \tag{D2}$$

From condition (D1), we have

$$\widetilde{L}_o^2 \leq \widetilde{L}_{b/o}^2 \frac{M_b}{M_o},$$

which implies that

$$\left(1 + 4\widetilde{L}_o^2\right)^{\frac{M_o}{M_b}} \leq \left(1 + 4\widetilde{L}_{b/o}^2 \frac{M_b}{M_o}\right)^{\frac{M_o}{M_b}},$$

and hence

$$\left(1 + 4\widetilde{L}_o^2\right)^{\frac{M_o}{M_b}} - \left(1 + 4\widetilde{L}_{b/o}^2\right) \leq \left(1 + 4\widetilde{L}_{b/o}^2 \frac{M_b}{M_o}\right)^{\frac{M_o}{M_b}} - \left(1 + 4\widetilde{L}_{b/o}^2\right) = g(\widetilde{L}_{b/o}^2), \tag{D3}$$

where the function $g$ is given by $g(x) = (1 + 4xM_b/M_o)^{M_o/M_b} - (1 + 4x)$. Taking the first derivative of $g$ gives

$$g'(x) = 4\left(\left(1 + 4x\frac{M_b}{M_o}\right)^{\frac{M_o}{M_b} - 1} - 1\right).$$

Since $M_o/M_b < 1$, for all $x \geq 0$, we have $(1 + 4xM_b/M_o)^{M_o/M_b - 1} < 1$ and hence $g'(x) < 0$; that is, $g$ is a decreasing function on $[0, +\infty)$. Consequently, $g(\widetilde{L}_{b/o}^2) \leq g(0) = 0$ and thus inequality (D3) implies that

$$\left(1 + 4\widetilde{L}_o^2\right)^{\frac{M_o}{M_b}} - \left(1 + 4\widetilde{L}_{b/o}^2\right) \leq 0,$$

or, equivalently,

$$\left(1 + 4\widetilde{L}_o^2\right)^{M_o} \leq \left(1 + 4\widetilde{L}_{b/o}^2\right)^{M_b}.$$

By using this inequality in Equation (D2), we obtain that

$$f(\widetilde{L}_o) = 1 + \frac{\sigma_b^2 \nu_b \widetilde{L}_{b/o}}{\sigma_o^2 \nu_o \widetilde{L}_o}.$$

In this expression, $f(\widetilde{L}_o)$ is inversely proportional to $\widetilde{L}_o$. Therefore, $f$ decreases on the interval $\left[0, \widetilde{L}_{b/o}\sqrt{M_b/M_o}\right]$. As a consequence, the minimum of $f(\widetilde{L}_o)$ must be located on the interval $\left[\widetilde{L}_{b/o}\sqrt{M_b/M_o}, +\infty\right)$. We now study the variations of $f$ on this interval, which means that condition (i) holds:

$$\widetilde{L}_o^2 M_o > \widetilde{L}_{b/o} M_b.$$

We assumed that condition (iii) holds when condition (i) is satisfied, which means that $f(\widetilde{L}_o)$ takes the form

$$f(\widetilde{L}_o) = 1 + \frac{\sigma_b^2 \nu_b \widetilde{L}_{b/o}}{\sigma_o^2 \nu_o \widetilde{L}_o} \left(\frac{\widetilde{L}_o^2}{M_b}\right)^{M_b} \left(\frac{M_o}{\widetilde{L}_{b/o}^2}\right)^{M_o} \left(\frac{M_b - M_o}{\widetilde{L}_o^2 - \widetilde{L}_{b/o}^2}\right)^{M_b - M_o}.$$

The derivative of $f(\widetilde{L}_o)$ can be expressed as

$$\frac{\partial f}{\partial \widetilde{L}_o}(\widetilde{L}_o) = \frac{\sigma_b^2 \nu_b \widetilde{L}_{b/o}}{\sigma_o^2 \nu_o \widetilde{L}_o^2 \left(\widetilde{L}_o^2 - \widetilde{L}_{b/o}^2\right)} \left(\frac{\widetilde{L}_o^2}{M_b}\right)^{M_b} \left(\frac{M_o}{\widetilde{L}_{b/o}^2}\right)^{M_o} \left(\frac{M_b - M_o}{\widetilde{L}_o^2 - \widetilde{L}_{b/o}^2}\right)^{M_b - M_o} \left[\widetilde{L}_o^2 (2M_o - 1) - \widetilde{L}_{b/o}^2 (2M_b - 1)\right]. \tag{D4}$$

If conditions (i) and (ii) are met, we have $\widetilde{L}_o^2 - \widetilde{L}_{b/o}^2 > 0$. Therefore, the stationary point for $f$ satisfies

$$\begin{aligned} \frac{\partial f}{\partial \widetilde{L}_o}(\widetilde{L}_o) = 0 &\Leftrightarrow \widetilde{L}_o^2 (2M_o - 1) = \widetilde{L}_{b/o}^2 (2M_b - 1) \\ &\Leftrightarrow \widetilde{L}_o = \widetilde{L}_{b/o}\sqrt{\frac{2M_b - 1}{2M_o - 1}}. \end{aligned} \tag{D5}$$

Since $M_o < M_b$, it follows that

$$\widetilde{L}_{b/o}\sqrt{\frac{2M_b - 1}{2M_o - 1}} > \widetilde{L}_{b/o}\sqrt{\frac{M_b}{M_o}}.$$

We are now interested in examining the behavior of $f$ on the intervals $\left[\widetilde{L}_{b/o}\sqrt{M_b/M_o}, \widetilde{L}_{b/o}\sqrt{(2M_b - 1)/(2M_o - 1)}\right]$ and $\left[\widetilde{L}_{b/o}\sqrt{(2M_b - 1)/(2M_o - 1)}, +\infty\right)$. For the first interval, we can show that $f$ is decreasing since $\partial f(\widetilde{L}_o)/\partial \widetilde{L}_o < 0$ from Equation (D4) if

$$\widetilde{L}_o^2 (2M_o - 1) < \widetilde{L}_{b/o}^2 (2M_b - 1) \;\Leftrightarrow\; \widetilde{L}_o < \widetilde{L}_{b/o}\sqrt{\frac{2M_b - 1}{2M_o - 1}}.$$

Similarly, for the second interval, we can show that $f$ is increasing since $\partial f(\widetilde{L}_o)/\partial \widetilde{L}_o > 0$ from Equation (D4) if

$$\widetilde{L}_o^2 (2M_o - 1) > \widetilde{L}_{b/o}^2 (2M_b - 1) \;\Leftrightarrow\; \widetilde{L}_o > \widetilde{L}_{b/o}\sqrt{\frac{2M_b - 1}{2M_o - 1}}.$$

Therefore, the stationary point (D5) is the unique minimum of $f$. Finally, multiplying both sides of Equation (D5) by $h_o\sqrt{2M_o - 1}$ yields Equation (26).