Environmental Data Science (2025), 4: e1, 1–11
doi:10.1017/eds.2024.47

METHODS PAPER

Scalable data assimilation with message passing

Oscar Key[1,*], So Takao[1,2,*], Daniel Giles[1,*] and Marc Peter Deisenroth[1,3]

[1]UCL Centre for Artificial Intelligence, University College London, London, United Kingdom
[2]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, United States
[3]The Alan Turing Institute, London, United Kingdom
Corresponding author: Oscar Key; Email: oscar.key.20@ucl.ac.uk

Received: 31 July 2024; **Accepted:** 28 September 2024

Keywords: Bayesian inference; data assimilation; distributed computation; message passing

Abstract

Data assimilation is a core component of numerical weather prediction systems. The large quantity of data processed during assimilation requires the computation to be distributed across increasingly many compute nodes; yet, existing approaches suffer from synchronization overhead in this setting. In this article, we exploit the formulation of data assimilation as a Bayesian inference problem and apply a message-passing algorithm to solve the spatial inference problem. Since message passing is inherently based on local computations, this approach lends itself to parallel and distributed computation. In combination with a GPU-accelerated implementation, we can scale the algorithm to very large grid sizes while retaining good accuracy and compute and memory requirements.

Impact Statement

This article addresses scalability issues with one of the core algorithms in numerical weather prediction systems. Solving these issues contributes to producing higher resolution and more frequently updated weather forecasts. Improved forecasts are an important tool for mitigating and adapting to climate change, with applications, such as predicting the output of wind and solar power, and warning about extreme weather events.

1. Introduction

Data assimilation (DA) is a core component of numerical weather prediction (NWP) systems. The goal of DA is to provide the best estimate of the current state of the dynamical system (the atmosphere in NWP). This is achieved by combining the forecasted state of the system with measurements, for example, satellites (tracks of which can be seen in Figure 1), sensors on the ground and weather balloons. A corrected forecast is then produced using the resultant estimate as the initial condition. Due to the number of observations and dimensionality of the state, DA consumes a large amount of computation time and memory.

In recent years, the scalability of DA approaches has been pushed to the limit, with operational weather centers launching programs to solve the problem (Bauer et al. 2020). Several challenges are identified.

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

*Equal contribution

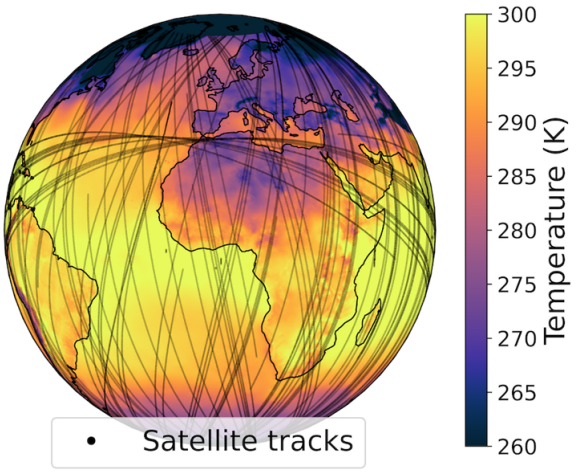

Figure 1. *Surface temperature computed by message passing from satellite observations. The lines show the locations of the observations.*

First, the amount of data that must be processed has grown dramatically due to the increasing availability of satellite observations and the higher resolution of forecasting models. Second, while the total computational power available continues to increase, this is no longer due to individual cores getting faster but instead in the form of expanded parallelism. Thus, DA algorithms must be able to take full advantage of parallel and distributed hardware.

4D-Var and three-dimensional (3D)-Var, and their derivatives, are the DA algorithms used by many systems for weather forecasting and other applications, such as ocean simulation (Bonavita and Lean 2021; Saulter et al. 2020). A standard approach to distribute these algorithms across many compute nodes is spatial parallelism using domain decomposition (D'Amore et al. 2014; Arcucci et al. 2015). The geographic area covered by the model is divided into overlapping subdomains; each subdomain is assigned to a single compute node that solves the assimilation problem in that subdomain. A key limitation of this approach is that the overlapping regions between subdomains must be carefully synchronized between nodes to ensure that the states computed on each node are physically consistent with each other. These synchronization steps introduce a bottleneck due to communication overheads, and this can be exacerbated by any computing load imbalance between the subdomains. Although there have been attempts to relax the level of synchronization required (Cipollone et al. 2020), ideally synchronization requirements would be removed altogether.

In this article, we propose an alternative approach to DA that is designed with distributed computation in mind. We exploit the formulation of DA as a Bayesian inference problem (Evensen et al. 2022), which allows us to apply tools from the large-scale Bayesian inference literature. In particular, we develop a method based on considering DA as inference in a Gaussian Markov random field (GMRF) and develop a message-passing algorithm to perform inference on this field. This approach naturally supports domain decomposition across multiple nodes *without* requiring overlapping regions and, therefore, no synchronization is required. Only a small amount of data must be communicated between subdomains, and this can be performed asynchronously with the computation on each node. In this initial work, we consider only two-dimensional spatial inference problems, rather than the 3D or 3D-with-time problems commonly seen in applications. However, the framework we develop here is designed to be extended to the full 3D-with-time case. The contributions of this work are as follows:

- We propose an approach for expressing the DA problem as a message-passing algorithm.
- We develop a GPU-accelerated implementation for maximum a posteriori (MAP) inference, which naturally supports distributed computation for very large domains.

https://doi.org/10.1017/eds.2024.47 Published online by Cambridge University Press

- We demonstrate the efficacy of our algorithm on surface temperature data and demonstrate that our approach is a viable DA technique.
- We also include a fast, GPU-accelerated implementation of 3D-Var as a very strong baseline.

We release our message passing and 3D-Var implementations, and code to reproduce our experiments, at github.com/oscarkey/message-passing-for-da.

2. Background

In this section, we recall the Bayesian formulation of DA and discuss several inference algorithms. In particular, we introduce the message-passing algorithm that we apply in this work.

2.1. Data assimilation as Bayesian inference

DA comprises a set of techniques in NWP that aim to combine earth observations with assumptions about the state of the weather to produce an updated estimate of the weather. In this work, we simplify the full DA problem to spatial inference of a single variable. Let $\boldsymbol{f} = (f_1, \ldots, f_n) \in \mathbb{R}^n$ be a weather variable, such as surface temperature, that is discretized on a large 2D spatial grid consisting of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. We also have m observations, $\boldsymbol{y} \in \mathbb{R}^m$, which may either be derived from remote sensing products or direct observations of atmospheric variables. From a probabilistic perspective, DA can then be understood as a Bayesian inference problem: our assumptions about the weather can be encoded as a prior $p(\boldsymbol{f})$, the observations as the likelihood $p(\boldsymbol{y}|\boldsymbol{f})$, and our updated estimate as the posterior computed via Bayes' theorem as $p(\boldsymbol{f}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f})$. n is typically in the billions and m in the tens of millions (MetOffice 2024), making direct inference using Bayes' theorem impossible. Thus, the choice of prior and likelihood is heavily influenced by the need to make inference efficient, and we discuss several options in the next section.

In operational DA systems, the problem is more complex, consisting of a 3D spatiotemporal grid and multiple weather variables at each grid point. However, inference in the simplified setting above is still challenging for large n and m. In Section 5, we discuss extensions to the full DA problem.

2.2. Existing methods for large-scale inference

2.2.1. Optimal interpolation

Optimal interpolation (Kalnay 2003, Section 5.4.1), is virtually synonymous with Gaussian process (GP) regression from the machine learning literature (Rasmussen and Williams 2006). This is the basis of all the DA methods discussed in this work, with subsequent methods being approximations to it. GP regression assumes a Gaussian prior $p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}_b,)$ and likelihood $p(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{H}\boldsymbol{f}, \boldsymbol{R})$. Here, $\boldsymbol{f}_b \in \mathbb{R}^n$ is the prior mean, and $\in \mathbb{R}^{n \times n}$ the prior covariance defined by a function k (known as a reproducing kernel), where $\boldsymbol{\Sigma}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. $\boldsymbol{H} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{R} \in \mathbb{R}^{m \times m}$ are the linear observation operator and diagonal error covariance, which are either determined using prior knowledge or learned from data. Under this prior and likelihood, the posterior $p(\boldsymbol{f}|\mathbf{y})$ is also a Gaussian, having a closed form expression. Unfortunately, computing the desired posterior costs $\mathcal{O}(l^3 + nl)$, where l is the number of grid points at which there are observations. A modern solution to large-scale inference with GPs is to use inducing-point methods, which approximate the observations with a much smaller number of pseudo data points (Titsias 2009). However, the estimates obtained may be too crude for practical use at the scale that is typically considered in numerical weather forecasting.

2.2.2. Gaussian Markov random fields

Another solution to reducing the cubic cost is to use a prior $p(\boldsymbol{f})$ defined by a GMRF (Rue and Held 2005). This approach makes use of the spatial interpretation of $\boldsymbol{f}$ to make a Markovian assumption that each f_i only directly depends on other f_j in its neighborhood. Additionally, the prior $p(\boldsymbol{f})$ and inference process are expressed in terms of the inverse of the covariance matrix, known as the precision. Under the Markovian assumption, the precision matrix is sparse, allowing inference in GMRFs to scale $\mathcal{O}(n^{3/2})$

https://doi.org/10.1017/eds.2024.47 Published online by Cambridge University Press

in 2D and $\mathcal{O}(n^2)$ in 3D (the complexity does not depend on the number of observations). In addition, the INLA (Integrated Nested Laplace Approximation) framework (Rue et al. 2009) makes it possible to handle nonlinear observation models and infer the model hyperparameters from data, although this requires further approximations. A downside of the approach is that it is inherently sequential. Thus, it cannot make effective use of modern parallel computing hardware, such as GPUs, or be distributed across several compute nodes.

2.2.3. 3D-Var

Another alternative that reduces the cost of GP regression is to only compute the MAP estimate, rather than the full posterior, that is find $\boldsymbol{f}_{\text{MAP}} = \arg\max_{\boldsymbol{f}} p(\boldsymbol{f}|\mathbf{y})$. This is the approach taken by 3D-Var, which is known as a "variational" method. The maximization can also be expressed as minimizing the cost function $J[\boldsymbol{f}] = -\log p(\boldsymbol{f}|\mathbf{y}) = -\log p(\mathbf{y}|\boldsymbol{f}) - \log p(\boldsymbol{f}) + C$, where C is a constant that does not depend on $\boldsymbol{f}$. Substituting in the GP prior and Gaussian likelihood, the cost function becomes

$$J[\boldsymbol{f}] = \frac{1}{2}(\mathbf{y} - \boldsymbol{H}(\boldsymbol{f}))^T \boldsymbol{R}^{-1}(\mathbf{y} - \boldsymbol{H}(\boldsymbol{f})) + \frac{1}{2}(\boldsymbol{f} - \boldsymbol{f}_b)^{T-1}(\boldsymbol{f} - \boldsymbol{f}_b). \tag{1}$$

In practice, this is minimized using an optimizer, for example, L-BFGS (Liu and Nocedal 1989) (as seen in D'Amore et al. (2015)) or a Krylov solver (as seen in Bauer et al. (2020)). 3D-Var is more flexible than optimal interpolation, as it can handle nonlinear observation models $\boldsymbol{H}$. However, it has the downside of not being able to provide uncertainty estimates, as it is a MAP estimator. In weather forecasting applications, 3D-Var is usually extended to 4D-Var, which includes a time component.

2.3. Factor graphs and inference with message passing

In this work, we perform inference using message passing (Kschischang et al. 2001), which computes the marginals of any joint probability model that can be expressed as a factor graph. We introduce message passing for a general continuous distribution $g(\boldsymbol{f}) = g(f_1, \ldots, f_n)$, and specialize it to our posterior in Section 3.2. $g(\boldsymbol{f})$ can be expressed as a factor graph if it has a known decomposition

$$g(\boldsymbol{f}) \propto \prod_{i=1}^{n} \phi_i(f_i) \prod_{j=1}^{n} \phi_{ij}(f_i, f_j), \tag{2}$$

where ϕ_i and $\phi_{i,j}$, referred to as the *nodewise* and *pairwise factors*, respectively, are functions from a variable, or a pair of variables, to $\mathbb{R}$. These do not have to be probability distributions. The factorization in (2) induces a sparse graph on the variables $\{f_i\}_{i=1}^n$, where two nodes f_i and f_j are connected if and only if $\phi_{i,j}$ is nonconstant. Given such a graph, the algorithm associates a pair of *messages* on each edge at each iteration t: a message $m_{ij}^t = \left(a_{ij}^t, b_{ij}^t\right) \in \mathbb{R} \times \mathbb{R}$ from f_i to f_j, and a similar message m_{ji}^t from f_j to f_i. We use the message-passing variant introduced by Ruozzi and Tatikonda (2013) (in turn a generalization of Wiegerinck and Heskes (2002)). This is summarized in Algorithm 1 and illustrated for our application in Figure 2, and Appendix A describes it in more detail.

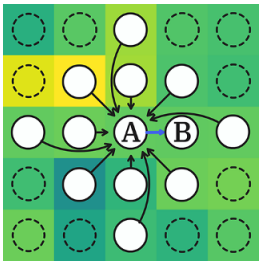

Figure 2. *Illustration of node A sending a message to node B. Black arrows indicate incoming messages that are combined to compute the outgoing message in blue.*

https://doi.org/10.1017/eds.2024.47 Published online by Cambridge University Press

Algorithm 1 Re-weighted message passing.

procedure MESSAGE PASSING ($\{f_i\}_{i=1}^n$, $\{\phi_i\}_{i=1}^n$, $\left\{\phi_{ij}\right\}_{i,j=1}^n$) ▷ input is factor graph
  $m_{ij}^{t=0} = \left(0, 10^{-8}\right)$ for all i,j
  for $t \in \{1, \ldots T\}$ **do**
    for f_i in the graph **do**
      for $f_j \in \left\{f_j \sim f_i\right\}$ **do**
        $m_{ij}^t = \text{COMPUTE OUTGOING MESSAGE}\left(\mathrm{c}, \phi_{\mathrm{i}}, \left\{\left(\phi_{ki}, \mathrm{m}_{ki}^{\mathrm{t}-1}\right) : \left\{\mathrm{f_k} \sim \mathrm{f_i}\right\} \setminus \mathrm{f_j}\right\}\right)$
      end for.
    end for.
  end for.
  return $g(f_i) = \text{COMPUTE MARGINAL}, \left\{\mathrm{m}_{ki}^{\mathrm{T}} : \mathrm{f_k} \sim \mathrm{f_i}\right\}$ for all i.
end procedure

We use the notation $\{f_k \sim f_i\}$ to denote the set of all variables f_k that are connected with f_i, COMPUTE OUTGOING MESSAGE() and COMPUTE MARGINAL() are defined formally in Appendix B, T is the total number of iterations, and $c \in \mathbb{R}$ is a hyperparameter to re-weight contributions of the messages, necessary to aid convergence. Note that each iteration of the inner for loops is independent, and each node only writes and reads messages with the nodes directly connected to it. The algorithm is therefore very amenable to distributed computation.

While this instance of message passing can theoretically compute both the mean and variance of the marginals, in practice, the variance estimates are biased and do not provide useful estimates of the uncertainty (Weiss and Freeman 1999). Thus, we only use message passing to compute the posterior mean, the MAP estimate of the posterior. This makes message passing an alternative to 3D/4D-Var.

3. DA with message passing

Our method begins by placing a Matérn GP prior over the domain, which is the de facto standard model choice in spatial geostatistics (Guttorp and Gneiting 2006), and assuming a Gaussian likelihood. We discretize the prior to a GMRF and derive the corresponding factor graph. Then, we apply a message-passing algorithm to the graph and the observations to compute the marginal posterior means.

3.1. Derivation of the factor graph

The Matérn GP prior can be characterized as the solution to a stochastic partial differential equation (SPDE) of the form

$$\left(\kappa^2 - \Delta\right)^{\alpha/2} f = \mathbf{W}, \tag{3}$$

where f is the process, Δ is the Laplacian operator, κ and α are the positive hyperparameters, and $\boldsymbol{\mathcal{W}}$ is the spatial white-noise process with spectral density $\sigma^2 q$, for hyperparameters $\sigma, q \in \mathbb{R}$. Following Lindgren et al. (2011), we first derive a GMRF representation of the Matérn GP by discretizing this SPDE using finite differences (finite elements would also be possible). On a uniform 2D grid with step sizes Δx and Δy in the x and y directions, respectively, this yields a random matrix–vector system

$$\boldsymbol{L}\boldsymbol{f} = \boldsymbol{w}, \quad \text{where} \quad \boldsymbol{w} = \sqrt{\frac{\sigma^2 q}{\Delta x \Delta y}}\mathbf{z}, \quad \mathbf{z} \in \mathbb{R}^n \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{I}_n). \tag{4}$$

Here, $\boldsymbol{L} \in \mathbb{R}^{n \times n}$ is a matrix representing the operator $\mathcal{L} := (\kappa^2 - \Delta)^{\alpha/2}$ under discretization, which is guaranteed to be sparse if the exponent $\alpha/2$ is an integer (Lindgren et al. 2011), and $\boldsymbol{f}, \boldsymbol{w}$ are

https://doi.org/10.1017/eds.2024.47 Published online by Cambridge University Press

finite-dimensional vector representations of the random fields f and $\mathcal{W}$ (see Appendix C for details on the discretization). Now, (4) implies that $\boldsymbol{f}$ is a Gaussian random variable of the form

$$\boldsymbol{f} \sim \mathcal{N}\left(\mathbf{0}, \left(\gamma \boldsymbol{L}^T \boldsymbol{L}\right)^{-1}\right), \quad \text{where} \quad \gamma := \frac{\Delta x \Delta y}{\sigma^2 q}, \tag{5}$$

which is a GMRF, since its precision matrix $\boldsymbol{P} := \gamma \boldsymbol{L}^T \boldsymbol{L}$ is sparse (here, $\boldsymbol{L}$ is sparse and banded). We can build a graph from this GMRF by the following simple rule: Take all of $f_1, \ldots, f_n$ as nodes in the graph and connect two nodes f_i and f_j by an edge if $[\boldsymbol{P}]_{ij} \neq 0$. Then, we have

$$p(\boldsymbol{f}) \propto \exp\left(-\frac{1}{2}\boldsymbol{f}^T \boldsymbol{P} \boldsymbol{f}\right) = \exp\left(-\sum_{i=1}^{n} \frac{1}{2}[\boldsymbol{P}]_{ii} f_i^2 - \sum_{j \sim i} [\boldsymbol{P}]_{ij} f_i f_j\right), \tag{6}$$

where $\Sigma_{j \sim i}(\cdot)$ denotes the sum over all indices j that are adjacent to i in the graph. Setting

$$\phi_i(f_i) := \exp\left(-\frac{1}{2}[\boldsymbol{P}]_{ii} f_i^2\right) \quad \text{and} \quad \phi_{ij}(f_i, f_j) := \exp\left(-[\boldsymbol{P}]_{ij} f_i f_j\right), \tag{7}$$

we have that the prior $p(\boldsymbol{f}) \propto \prod_{i=1}^{N} \phi_i(f_i) \prod_{j \sim i} \phi_{ij}(f_i, f_j)$, thus we have a factor graph representation.

3.2. Computing the posterior mean with message passing

Having calculated the factor graph corresponding to the prior, we can now apply message passing (Algorithm 1) to combine this with the observations and compute the posterior marginals. We make several modifications to tailor the algorithm to DA, which we summaries below and detail in Appendix B.

3.2.1. Including observations

To add information about the observations $\mathbf{y}$ into our factor graph, we modify the nodewise factor in (7) by $\phi_i(f_i) \mapsto p(y_i|f_i)\phi_i(f_i)$. In our experiments, we assume that the weather variable is noisely observed at a subset of grid cells, where the noise $\sigma_y^2 \in \mathbb{R}$ is constant for all observations. Thus, we set $p(y_i|f_i) = \mathcal{N}\left(y_i|f_i, \sigma_y^2\right)$ at grid points $\mathbf{x}_i$ where there is an observation, and $p(y_i|f_i) = \mathcal{N}(y_i|0, z)$, for very large z, where there is not.

3.2.2. Update damping

To improve the stability of the algorithm, we follow Pretti (2005) and dampen the updates of the messages, replacing line 6 of Algorithm 1 with

$$m_{ij}^{t+1} = (1-\eta)m^t + \eta \textsc{Compute Outgoing Message}(\cdot),$$

where $\eta \in (0,1)$ is a hyperparameter, which we refer to as the learning rate.

3.2.3. Early stopping

To avoid specifying the total number of iterations as a hyperparameter, we choose a generic large T and stop when the change in message between iterations is smaller than a threshold.

3.2.4. Multigrid

We apply a multigrid technique to speed up the convergence of the algorithm. These have been used extensively when solving partial differential equations numerically, and provide an efficient way of accelerating the convergence of iterative approaches if multiscale phenomena are being modeled, with grids at different resolutions capturing different spatial scales. In the case of message passing, we can intuitively view information propagating from the observation locations across the graph. If the density of the observations is low, this can take many iterations. To solve this, the multigrid approach starts with a

https://doi.org/10.1017/eds.2024.47 Published online by Cambridge University Press

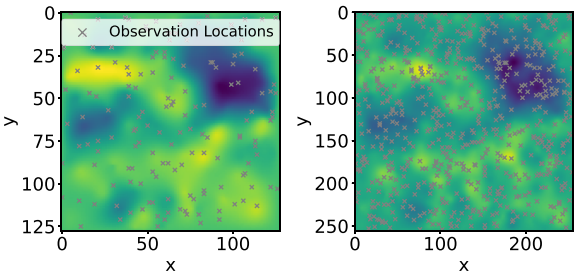

Figure 3. *Illustration of the multigrid implementation, showing the marginal means computed at two levels (*128×128 *to* 256×256*) of resolution on the simulated data.*

low-resolution grid, and iterates message passing until convergence—this is fast on a low-resolution grid. It then doubles the size of the grid, initializing the messages to the converged messages from the previous grid. This process is repeated until we reach the target grid size. Observations are taken on the target resolution and introduced at each multigrid level when the observation location exactly coincides with the grid level coordinates. Figure 3 illustrates the procedure.

3.2.1. Computational efficiency

The Markovian assumption made by the GMRF prior from which we derive the factor graph results in a sparse graph in which each node is only connected to nodes in its local area. The connectivity is also very regular, with each node connected to the same set of relative nodes except at the edges of the graph. These two properties make the graph high amenable to GPU computation, using an approach similar to (Zhou et al. 2022) but optimized for our particular graph structure.

The main advantage of the message-passing approach is that it can naturally be distributed by dividing the nodes of the graph into subdomains and splitting them between compute nodes. The only communication required between compute nodes is to update the messages on the borders of the subdomains, which is a small amount of data. Additionally, there is no requirement that the border messages are updated every iteration, so they could be updated asynchronously to the computation within each subdomain.

4. Experiments

We evaluate the performance of message passing on both simulated data and a more realistic surface temperature DA problem. We compare against two baselines: the GMRF method, using the R-INLA implementation (Lindgren and Rue 2015), and 3D-Var. Note that R-INLA computes the exact marginals of the posterior, while 3D-Var and message passing are approximate methods that compute only the mean. Thus, R-INLA provides a reference for the best-case error that the approximate methods could achieve.

GPU-accelerated 3D-Var implementation. The 3D-Var cost function is obtained by substituting in the same GMRF prior and Gaussian likelihood as used for message passing, and then minimized using the L-BFGS optimizer. We use our own GPU-accelerated implementation using the experimental support for sparse linear algebra in JAX (Bradbury et al. 2018) and JAXopt (Blondel et al. 2021). We expect this implementation to be significantly faster than the CPU implementations currently in deployment. We also implement message passing in JAX with GPU acceleration, while R-INLA runs on the CPU only.

Hyperparameters. We perform a grid search, reported in Appendix D.1, to select the message weighting and learning rate hyperparameters of message passing, and the early stopping thresholds for both message passing and 3D-Var. We note that selecting the message weighting and learning rate is quite easy: the algorithm converges for a wide range of choices and, if it does converge, the exact choice has only a small effect on the speed of convergence. Appendix E gives the remaining details of all experiments.

https://doi.org/10.1017/eds.2024.47 Published online by Cambridge University Press

Table 1. *Comparison on simulated data. We give the mean over three ground truths; we do not observe significant variance (therefore omitted). Bold indicates where either 3D-Var or message passing performed better. R-INLA is included to show the minimal achievable error given the prior, as it computes an exact posterior*

		RMSE			duration (seconds)		
grid size	observations	R-INLA	3D-Var	MP	R-INLA	3D-Var	MP
256×256	1%	0.192	**0.202**	0.213	19.4	**3.6**	7.6
	5%	0.093	0.093	0.093	21.6	4.1	**1.7**
	10%	0.069	0.069	**0.068**	22.5	4.5	**0.9**
512×512	1%	0.101	0.128	**0.127**	107.9	**3.9**	15.9
	5%	0.047	0.048	0.048	104.8	5.3	**4.2**
	10%	0.034	0.036	**0.034**	99.8	5.9	**2.5**
1024×1024	1%	0.050	0.116	**0.066**	601.6	**4.5**	50.5
	5%	0.024	0.026	**0.024**	848.7	**8.0**	13.8
	10%	0.017	0.020	**0.017**	547.5	11.5	**8.7**

4.1. Effect of domain size and observation density

Our first set of experiments are performed on simulated data. We sample a square ground truth field from a GMRF prior, randomly select a given fraction of the grid points as observations, and perform inference against these observations and the same prior. Table 1 shows the RMSE between the posterior mean and the ground truth and the time taken, as we vary the grid size and the observation density. In the message-passing runs, the multigrid approach is used, with a base grid size of 32×32 in all cases.

The results show that 3D-Var and message passing achieve similar RMSEs in all cases, although both have more error than R-INLA when only 1% of the grid is observed. When 5% and 10% of the grid is observed, 3D-Var and message passing take similar amounts of time; however, message passing is significantly slower for a larger grid when only 1% is observed. While the iteration time of message passing is independent of the number of observations, as the observation density falls it takes an increasing number of iterations for the information to propagate from the observed points across the grid, thus early stopping happens later. R-INLA is much slower than the other methods, because it is a sequential method that cannot take advantage of the GPU.

4.2. Large-scale example

For a more realistic use case, we consider the global surface temperature field. We take the ground truth data from a run of the Met Office's Unified Model (Walters et al. 2019) at N1280 resolution, where the data are valid for 06UTC 2020-01-01. To avoid issues with boundary conditions, we consider a clipped domain with dimensions $2500 \times 1500 = 3.75M$ grid points. We use spherical polar coordinates. The observation locations are generated from the geographical positions (latitude, longitude) of weather-focused satellites calculated over a 3-hour window, which corresponds to $\approx 8\%$ of the grid being observed. As the prior mean, we select a climatology mean of the global surface temperature calculated from ERA5 (Hersbach et al. 2020).

Figure 1 highlights the resultant mean estimates from the message-passing approach, while error plots are shown in Figure 7 in Appendix D. 3D-Var achieved an area-weighted RMSE of 2.33 K and took 16 seconds, while message passing achieved an area-weighted RMSE of 1.23 K and took 115 seconds. For comparison, the area-weighted RMSE calculated for the prior mean (ERA5) against the high-resolution

https://doi.org/10.1017/eds.2024.47 Published online by Cambridge University Press

temperature field is 2.78 K. R-INLA did not complete after >1 hour of processing time; thus, we do not include its results.

5. Conclusion

In this article, we present a new perspective on DA based on insights from the literature on large-scale Bayesian inference. We demonstrate that our message-passing approach is viable and can, in many scenarios, produce results competitive with a GPU-accelerated 3D-Var implementation. In the era of large heterogeneous computing systems, the scalability issues with state-of-the-art variational DA are well documented and the design of the proposed method should offer improved scalability. However, further research is required to determine if this design offers advantages for operational-scale problems.

We make several simplifying assumptions. First, we have only considered spatial inference in two dimensions. However, our approach can be extended to full spatiotemporal inference as we describe in Appendix C.3. We have also only considered a single weather variable and linear observation operators. However, this can be extended relatively easily to multiple weather variables under the current framework, by assuming that they are independent under the prior. Further work is required for variables coupled in the prior. It may also be possible to support nonlinear observation operators using iterative linearization techniques (Kamthe et al. 2022).

The primary limitation of our approach is that message passing only reliably computes the posterior mean; the obtained posterior marginal uncertainties are inherently biased when the graph is loopy (Weiss and Freeman 1999). Thus, we cannot get an accurate estimate of the uncertainty in the assimilated state. It also requires Gaussianity assumption in the prior, although for non-Gaussian priors arising from nonlinear stochastic PDEs, we may be able to handle this using an iterative linearization method (Anderka et al. 2024). Finally, due to our unreliable uncertainty estimates, we cannot make use of the marginal likelihood to learn the hyperparameters of the prior, for example, the kernel lengthscale. Currently, we handle this using cross-validation on held-out observations. We note, however, that all of these limitations are shared with popular large-scale DA methods, such as 3D-Var and resolving these issues will be a significant step forward for future DA research.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2024.47.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/eds.2024.47.

Acknowledgments. The authors are grateful for the technical assistance of Dr. Cyril Morcrette for providing the Met Office's Unified Model high-resolution temperature data. The authors are also grateful for Prof. Mohammad Emtiyaz Khan and Prof. Nicholas Ruozzi for useful discussions and feedbacks.

Author contribution. Conceptualization: ST and MD. Methodology: ST and MD. Software: OK and DG. Data curation: DG. Data visualization: DG and OK. Writing—original draft: OK. Writing—review and editing: OK, ST, DG, and MD. Supervision: ST and MD. All authors approved the final submitted draft.

Data availability statement. Our message passing and 3D-Var implementations, and code to reproduce our experiments, are available at https://github.com/oscarkey/message-passing-for-da. It is also archived with at https://doi.org/10.5281/zenodo.14176688. The data for the surface temperature data experiments are taken from the Met Office's Unified Model, and thus sadly cannot be publicly released. The data for the other experiments are generated automatically by the experiments.

Provenance. This article was accepted into the Climate Informatics 2024 (CI2024) Conference. It has been published in Environmental Data Science on the strength of the CI2024 review process.

Funding statement. OK acknowledges support from the Engineering and Physical Sciences Research Council with grant number EP/S021566/1. ST is supported by a Department of Defense Vannevar Bush Faculty Fellowship held by Prof. Andrew Stuart, and by the SciAI Center, funded by the Office of Naval Research (ONR), under Grant Number N00014-23-1-2729.

Competing interest. The authors declare none.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Anderka R, **Deisenroth MP, and Takao S** (2024) Iterated INLA for state and parameter estimation in nonlinear dynamical systems. In *Conference on Uncertainty in Artificial Intelligence.*

Arcucci R, **D'Amore L**, **Celestino S**, **Scotti G and Laccetti G** (2015) A parallel approach for 3D-variational data assimilation on GPUs in ocean circulation models. *International Journal of Computer and Information Engineering 9*(5), 1204–1210.

Bauer P, **Quintino T**,**Wedi N**, **Bonanni A**, **Chrust M**, **Deconinck W**, **Diamantakis M**, **Düben P**, **English S**, **Flemming J**, **Gillies P**, **Hadade I**, **Hawkes J**, **Hawkins M**, **Iffrig O**, **Kühnlein C**, **Lange M**, **Lean P**, **Marsden O**, **Müller A**, **Saarinen S**, **Sarmany D**, **Sleigh M**, **Smart S**, **Smolarkiewicz P**, **Thiemert D**, **Tumolo G**, **Weihrauch C**, **Zanna C, and Maciel P** (2020). *The ECMWF Scalability Programme: Progress and Plans.* eng. URL: https://www.ecmwf.int/node/19380.

Blondel M, **Berthet Q**, **Cuturi M**, **Frostig R**, **Hoyer S**, **Llinares-López F**, **Pedregosa F, and Vert J-P** (2021) "Efficient and modular implicit differentiation." arXiv: 2105.15183.

Bonavita M and Lean P (2021) 4D-Var for numerical weather prediction. *Weather 76*(2), 65–66.

Bradbury J, **Frostig R**, **Hawkins P**, **Johnson MJ**, **Leary C**, **Maclaurin D**, **Necula G**, **Paszke A**, **VanderPlas J**, **Wanderman-Milne S, and Zhang Q** (2018). *JAX: Composable Transformations of Python+NumPy Programs.* Version 0.3.13. URL: http://github.com/google/jax.

Cipollone A, **Storto A and Masina S** (2020) Implementing a parallel version of a Variational scheme in a global assimilation system at Eddy-resolving resolution. *Journal of Atmospheric and Oceanic Technology 37*(10), 1865–1876.

D'Amore L, **Arcucci R**, **Carracciuolo L and Murli A** (2014) A scalable approach for Variational data assimilation. *Journal of Scientific Computation 61*(2), 239–257.

D'Amore L, **Laccetti G**, **Romano D**, **Scotti G and Murli A** (2015) Towards a parallel component in a GPU–CUDA environment: A case study with the L-BFGS Harwell routine. *International Journal of Computer Mathematics 92*(1), 59–76.

Evensen, G., **Vossepoel, F. C., and van Leeuwen, P. J.** (2022). *Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem.*

Guttorp P and Gneiting T (2006) Studies in the history of probability and statistics XLIX on the Matern correlation family. *Biometrika 93*(4), 989–995.

Hersbach H, **Bell B**, **Berrisford P**, **Hirahara S**, **Horányi A**, **Muñoz-Sabater J**, **Nicolas J**, **Peubey C**, **Radu R**, **Schepers D**, **Simmons A**, **Soci C**, **Abdalla S**, **Abellan X**, **Balsamo G**, **Bechtold P**, **Biavati G**, **Bidlot J**, **Bonavita M**, **De Chiara G**, **Dahlgren P**, **Dee D**, **Diamantakis M**, **Dragani R**, **Flemming J**, **Forbes R**, **Fuentes M**, **Geer A**, **Haimberger L**, **Healy S**, **Hogan RJ**, **Hólm E**, **Janisková M**, **Keeley S**, **Laloyaux P**, **Lopez P**, **Lupu C**, **Radnoti G**, **Rosnay P d**, **Rozum I**, **Vamborg F**, **Villaume S and Thépaut J-N** (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society 146*(730), 1999–2049.

Kalnay E (2003) *Atmospheric Modeling*. Data Assimilation and Predictability.

Kamthe S, **Takao S**, **Mohamed S, and Deisenroth M** (2022) Iterative state estimation in non-linear dynamical systems using approximate expectation propagation. In *Transactions on Machine Learning Research.*

Kschischang FR, **Frey BJ and Loeliger H-A** (2001) Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory 47*, 498–519.

Lindgren F and Rue H (2015) Bayesian spatial modelling with R-INLA. *Journal of Statistical Software 63*(19).

Lindgren F, **Rue H and Lindström J** (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology 73*(4), 423–498.

Liu DC and Nocedal J (1989). On the limited memory BFGS method for large scale optimization. In *Mathematical Programming* 45, 1–3, pp. 503–528.

MetOffice (2024) *Data Assimilation Methods.* URL: https://www.metoffice.gov.uk/research/weather/satellite-and-surfaceassimilation/data-assimilation-methods (visited on 01/28/2024).

Pretti M (2005) A message-passing algorithm with damping. *Journal of Statistical Mechanics: Theory and Experiment 2005*(11), P11008.

Rasmussen CE and Williams CKI (2006) *Gaussian Processes for Machine Learning.*

Rue H and Held L (2005) *Gaussian Markov Random Fields: Theory and Applications.*

Rue H, **Martino S and Chopin N** (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology 71*(2), 319–392.

Ruozzi N and Tatikonda S (2013) Message-passing algorithms for quadratic minimization. *Journal of Machine Learning Research 14*(69), 2287–2314.

Saulter A, **Bunney C**, **King R and W J** (2020) An application of NEMOVAR for regional wave model data assimilation. *Frontiers in Marine Science 7*(579834).

Titsias M (2009) Variational learning of inducing variables in sparse gaussian processes. In: *Artificial Intelligence and Statistics*, pp. 567–574.

Walters D, **Baran AJ**, **Boutle I**, **Brooks M**, **Earnshaw P**, **Edwards J**, **Furtado K**, **Hill P**, **Lock A**, **Manners J**, **Morcrette C**, **Mulcahy J**, **Sanchez C**, **Smith C**, **Stratton R**, **Tennant W**, **Tomassini L**, **Van Weverberg K**, **Vosper S**, **Willett M**, **Browse J**, **Bushell A**, **Carslaw K**, **Dalvi M**, **Essery R**, **Gedney N**, **Hardiman S**, **Johnson B**, **Johnson C**, **Jones A**, **Jones C**, **Mann G**,

https://doi.org/10.1017/eds.2024.47 Published online by Cambridge University Press

Milton S, **Rumbold H**, **Sellar A**, **Ujiie M**, **Whitall M**, **Williams K and Zerroukat M** (2019) The met Office unified model global atmosphere 7.0/7.1 and JULES global land 7.0 configurations. *Geoscientific Model Development 12*(5), 1909–1963.

Weiss Y and Freeman W (1999) Correctness of belief propagation in gaussian graphical models of arbitrary topology. In *Advances in Neural Information Processing Systems.*

Wiegerinck W and Heskes T (2002) Fractional belief propagation. In *Advances in Neural Information Processing Systems 15.*

Zhou G, **Dedieu A**, **Kumar N**, **Lázaro–Gredilla M**, **Kushagra S, and George D** (2022) PGMax: Factor graphs for discrete probabilistic graphical models and loopy belief propagation in JAX. arXiv: 2202.04110.

Cite this article: Key O, Takao S, Giles D and Deisenroth MP (2025). Scalable data assimilation with message passing. *Environmental Data Science*, 4: e1. doi:10.1017/eds.2024.47