Response to Reviewer #1 (Reviewer's original text is in blue italics and author response is in black.)

Review of "The Module and Integrated Data Assimilation System at Environment and Climate Change Canada (MIDAS v3.9.1) by M. Buehner et al.

The manuscript provides an overview of the data assimilation software "MIDAS". The available data assimilation algorithms and observation types that can be assimilated are shortly described. In addition a short overview of further functionality like for processing ensembles, e.g. inflation schemes for the data assimilation, for observation pre-processing and estimation of observation impacts, analysis error estimation, and diagnostics and statistics is provided. The structure of the software with respect of the different programs that are included is described and examples for the modular structure are given. Likewise the parallelization strategy is shortly described. After these descriptions of the functionality and structure of the software a large number of application examples are provided (the manuscript states 'all applications ... are briefly descried". Partly, a figure is shown for an application but for some applications there only a short paragraph of text. The manuscript completes with a summary and short description of future development plans. There are no explicit conclusions, but a statement that the flexibility of MIDAS could make this software useful to data assimilation research at Canadian universities and the preliminary efforts have started to make the software accessible (lines 597-599).

We thank the reviewer for carefully reading our submission and providing extensive and helpful comments that should lead to substantial improvements.

The manuscripts fits into the scope of GMD, even though the software that is described is not a 'model' but a data assimilation software. From my own experience I know that it can be difficult to publish about software and finding a particular focus might be particularly difficult. Unfortunately, the authors here apparently tried to discuss too many aspects and missed to provide a sufficient focus.

We appreciate the reviewer sharing their experience concerning the difficulty in publishing about software. Since most existing scientific publications related to data assimilation for operational NWP include little high-level description of the software used, it is difficult for software developers to learn from the ideas and approaches used at other NWP centers. Anticipating potential difficulties in publishing a description of the MIDAS software, we submitted this manuscript only after receiving encouragement from one of the editors who assured us that such a manuscript could be suitable for GMD. We hope that, after making many improvements to the manuscript based on the reviewers' comments, this will be possible.

While the title and abstract state that the manuscript is about a particular version of MIDAS, the manuscript does not focus on this version and also contains parts in which parts of the development history are described. For a particular version of the software, it is however irrelevant if some feature was 'recently introduced'.

Since the MIDAS software is used both for operational prediction applications and for performing DA research by many ECCC scientists, it is constantly evolving. This means that the functionality included in MIDAS and even the way some existing functionality is implemented is changing through time. Therefore, to avoid confusion when describing the software functionality and its implementation, the manuscript focuses on a single version. The chosen version is the one used for the most recent upgrade of all of the ECCC operational prediction systems that took place during June 2024. This point is now made clearer in the revised manuscript.

Text added to abstract: The described version of MIDAS is part of the Canadian prediction systems that became operational in June 2024.
Text added to introduction: As the MIDAS software is continually being developed and modified to facilitate both research and future operational applications, it was decided to focus the description of functionality, software design, and implementation details on this particular version to avoid confusion.

The description of the functionality is very superficial. Many features, e.g. variants of supported covariance matrices (lines 155-166) or the variants of the ensemble-based Kalman filters LETKF (lines 182-189) are just shortly listed. Only, the variational DA algorithm is explained by equations. However, this looks unbalanced compared to the otherwise short descriptions by text (even more the equations seem to show a standard 3D-Var scheme with control vector transformation - for specialists from data assimilation this is just standard, while for non-specialists the description is too short).

The description of the functionality will be modified in the revised manuscript to provide more balance in the level of detail. The description of the basic variational approach will be reduced by removing the equation of the gradient and we will more clearly explain that the basic cost function is presented as a way of introducing some of the mathematical notation that is used later in the text. In other subsections where we choose to provide less detail due to existing publications that contain suitable explanations, the text will be modified to give more information on where the additional details can be found, including section and equation numbers from those publications.

Text modified in section 3:

[Text added to section 3.1:] The formulation of the cost function in MIDAS is presented to introduce the mathematical notation used later and to complement the discussion on software design regarding the implementation of these mathematical operations. [Equation for cost function gradient and some surrounding text was removed.] [Following text was removed from section 3.1:] Previously, only the stochastic formulation of the EnKF algorithm (Houtekamer et al., 2014) implemented independently of MIDAS was available.

The description of the modular software design also contains little details beyond that there are modules with different purposes (some defining entities, while other focus on operations) and that the software uses a structure similar to 'classes' in object-oriented programming (but it is not using classes, which are supported by Fortran).

To keep the overall length of the manuscript manageable, the description of the software design must necessarily be limited to a relatively high-level presentation. Since the target audience includes meteorologists and data assimilation scientists working at operational NWP centres that may not have extensive software development training, we chose to focus on the general guiding principles behind the software design and the advantages of following such an approach. The goal is to provide an informative example of the ideas and approaches taken at one NWP centre that could assist developers of similar data assimilation software at other operational NWP centres. The goal of the software design discussion is now better explained in the revised manuscript.

Text added at the beginning of section 4: This section provides a brief and high-level description of the MIDAS software design to provide an informative example that may be useful for developers of similar DA software at other operational NWP centres.

Likewise the parallelization strategy is contains very little details since it's also mainly text. The motivation and performance of the parallelization choice is not discussed.

The section describing the parallelization strategy provides only a brief description of all of the different ways data are distributed over MPI tasks within MIDAS. Since there are many different MPI distributions, we focus on simply describing each one and provide some explanation for where they are used within the different data assimilation algorithms. However, based on the reviewer’s comment the revised manuscript now provides more information, in places where it was not already given, on the motivation of why each approach is used in each case. Also an example of the execution time for reading and transposing the 4D ensemble in the context of the operational LETKF is provided.

Text added in section 5:
[Regarding the latitude-longitude distribution of gridded state vectors:] This choice of distribution facilitates several procedures related to the assimilation algorithms, including: calculating mean of ensemble, applying LETKF weights to ensemble perturbations, computing 4D-EnVar increments from localized ensemble.
[Regarding the execution time of reading the ensemble:] The entire process of reading and transposing the hourly 256-member ensemble takes approximately one minute of the total execution time of about 11 minutes in the current global operational LETKF (described in section 6.1).
[Regarding the random distribution of observations:] ...to ensure a nearly even computational load across MPI tasks when applying the observation operators

The application examples are short sections of 'all' possible applications. These are also not aimed at actually presenting what users of MIDAS could achieve, but seem to be rather aiming for showing the different operational or research applications at Environment and Climate Change Canada.

Regarding the choice of applications that are described, the conscious choice was made to only describe those MIDAS applications included in the most recent version of the ECCC operational prediction systems. These applications are rigorously tested and evaluated as part of the normal process when modifying operational prediction systems. The MIDAS software includes a wide range of functionality, much of it implemented to facilitate DA research on new approaches and methods that may not be fully evaluated yet. Therefore, to maintain a reasonable length for the “applications” section (and of the overall manuscript) it was decided to only describe the operational applications. This is made clearer in the revised manuscript.

Text added at the beginning of section 6: While MIDAS could be used for numerous other types of applications, it was chosen to only present those applications that have been rigorously tested and evaluated prior to their operational implementation.

The different sections are also too short to provide sufficient details so that one hardly learns about particular functionalities of MIDAS. Only sometime a reference to a more detailed publication is provided here.

We understand the reviewer’s overall concern that many of the sections do not contain enough information to fully understand each aspect of MIDAS. Being intended as the first publication focused on the MIDAS software, we have attempted to provide an overview of many aspects. It would not be possible to provide a much more detail for all of these aspects within a single publication while keeping to a reasonable overall length. Some existing publications cited in the manuscript describe specific scientific

aspects related to MIDAS functionality in more detail and it is planned to prepare future publications that will focus on other aspects, such as the computational efficiency of the MIDAS implementation of the LETKF algorithms.

Overall, the purpose of the manuscript is not clear. The abstract states that "The ... MIDAS software (version 3.9.1) is described...". However, while the authors also made this version of MIDAS available on github.com, the software does not seem to be intended to be used by other users. E.g., the documentation is mainly generated from in-code comments. It fails to provide sufficient information on how to use the software (e.g. structure of input files and configuration files). Obviously, just making a software available online and generating documentation from in-code comments is insufficient to enable other to use it. The manuscript is over all not at a sufficient level for a scientific publication, but it leaves the impression of a technical report. For a scientific publication it is far too superficial in the methodological and functionality sections, but also in the cursory descriptions of the applications. To this end, I don't see a chance to revise the manuscript to a sufficient level of a scientific publication, since this would essentially imply a full re-write. Accordingly, I can only recommend to reject the manuscript.

We thank the reviewer for pointing out the lack of clarity regarding the purpose and intended audience of the manuscript. This is now made much clearer in the revised manuscript. The intended purpose of the publication is indeed not to make the software usable by people outside of our organisation. Instead, the purpose is to target an audience of scientists and software developers at other operational NWP centers involved in developing similar DA software for both NWP and related Earth system components. By describing the overall design of MIDAS and the reasons behind the design choices it is hoped that such a publication would be helpful for this community as they consider how to improve their own software. Based on informal personal communications with such people during international conferences and workshops, we believe there would indeed be readers who could benefit from this description of the MIDAS software.

Text added near the end of the Introduction: This description of MIDAS is intended to be helpful for an audience of developers of similar DA software at other operational NWP centres. It provides one example of how such software can be designed and provides the basic rationale supporting the design choices. Since at this time MIDAS cannot be easily installed and used outside of ECCC, the goal is not to make MIDAS available to the greater DA community.

Apart from the recommendation above, a scientific publication about MIDAS could certainly be valuable if it is prepared with sufficient care. To this end, I like to provide some recommendations for a possible paper publication about MIDAS:

- Please be clear about the purpose of such a publication. It should contain clear explanations of sufficient depth and detail of the particularities. The aim should obviously be to make the software 'usable' by a wider group of readers - otherwise there is no point in publishing about it. (There are a number of articles published in GMD that discuss data assimilation software aimed at a wider group of possible users which might give indications of what can be successful in GMD)

We mention above the intended purpose of the manuscript, which will be made clearer in the revised manuscript. Unfortunately, we do not currently have the resources or mandate to make the MIDAS software usable by users outside of ECCC. However, as explained above, we still believe that there is a benefit for readers of GMD to publish this type of manuscript.

- Also please consider carefully the target audience of the publication. A wide audience would be useful to enhance the value of the publication

We mention above the intended target audience, which is now made clearer in the revised manuscript.

- For a scientific publication I recommend to avoid 'storytelling'. This occurs in various places of the manuscript. To give just one example, lines 86-95 state "The implementation of the 4D-EnVar algorithm in the existing 4D-Var software was not practical..." It is unlikely that this contains any useful information for the readers. Even more the information is irrelevant for version 3.9.1 of MIDAS.

We feel that a short description of the overall development path of MIDAS and the reasoning behind the choices that were made along the way can be useful for the intended audience of DA software developers at other operational NWP centers. Some world-class NWP centers have chosen to develop new DA software nearly “from scratch” and point to the poor design of their current operational software to justify this choice. Because of this, we think it is useful to provide a concrete example of an operational NWP center that chose a different strategy and the reasons behind this choice. As described in the manuscript, our strategy involved starting with the existing operational software and gradually refactoring the code to improve the overall design and to allow new functionality to be more easily implemented. The choice of either of these development strategies can have significant impacts on an operational NWP center and many factors need to be considered when choosing the best strategy for a particular NWP center. We hope our publication could contribute to this. Elements of this reasoning behind the description of the development strategy used for developing

MIDAS are added to the revised manuscript.

Text added to the beginning of section 2: A brief description of the initial development steps of MIDAS and the reasoning behind the chosen development strategy is provided as an example for DA software developers and scientists at other NWP centres of how the constraints and requirements shared among many such centres can be addressed.

- Useful would also be to be clear about the question 'Why this code version?'. The manuscript states that it is about 'v3.9.1'. As a sub-sub-version this seems to be arbitrary and it does not show an ambition of making the software usable (and useful) for others. E.g. one could introduce a new major release (with sufficiently major changes) and the intention of real open-source software. One could take this as the motivation to publish about it (obviously one cannot publish about each new release, but only the most relevant ones)

We chose to describe the MIDAS version used in the latest operational upgrade of all operational prediction systems at ECCC. This is the first version of the ECCC operational systems that uses MIDAS for applications other than only atmospheric DA. This is now made more clear in the revised manuscript as mentioned above.

- Application examples should be carefully chosen with the aim that they are relevant for the readers to learn from them.

We chose to describe all applications used in the operational ECCC prediction systems after the latest upgrade to these system (which took place in June 2024). Such upgrades only occur every 2-3 years and therefore mark a major milestone that culminates from many years of research and testing. This seems to us to be a logical way of choosing the applications to include. This is now mentioned in the revised manuscript, as mentioned above.

- To be useful for readers, the software has to be 'usable'. This implies a sufficient documentation of how it can be used. Also required would be example cases, e.g. toy models. Without this, there is little purpose of making the code available on Github. Achieving a sufficient level of usability is in fact a larger task. Here, one should also be carefully considering the question "Can we support users?" - if the answer is not clearly 'yes', one should perhaps refrain from publishing it open-source.

As mentioned above, the purpose of this manuscript is not to make the code available for general use. Instead it is to provide information about the functionality and design of MIDAS which we hope could be useful for the community of DA software developers at other NWP centers. By making the Fortran code itself available on Github, such

developers can see in concrete terms how we chose to implement certain aspects of our system. It is hoped that this could be helpful to others when deciding how to develop or improve their own software. We also note that, at least for operational NWP, most centers currently use DA software that they have developed themselves. In a few specific cases, software (both for DA and the forecast model) has been shared between the NWP centers of different countries (e.g. UK Met Office, Australia, South Korea), but this is still relatively rare and requires extensive negotiation and formal agreements between the organisations.

Text added in section 4: ...(primarily for information purposes, as no support is available for users external to ECCC, a public copy is maintained at...

As a final recommendation, I like to suggest to be particularly careful when comparing the functionality of the 'own' software to other existing software. Here the risk is high that one misses some functionality so that the description of the other software is incorrect or incomplete. For example in the manuscript it is stated, relating to e.g. the DART and PDAF data assimilation software, that 'other systems previously mentioned all need to be compiled and executed together with the forecast model and exchange information between DA and forecast model through subroutine' (lines 70-72). However, DART was initially designed to use only files-based transfer of data and a separate program for the assimilation. In contrast, since many years PDAF supports both the direct coupling into a model code and the separate assimilation program with file-based data exchange (see e.g. https://pdaf.awi.de/trac/wiki/GeneralImplementationConcept?version=14). Further, the statement "While DART and PDAF were developed exclusively for applying ensemble DA algorithms to many different applications" (lines 64-65) is incorrect for PDAF. While PDAF was not 'primarily developed for operational NWP' (line 66), PDAF is applied in both research and operational applications. E.g. PDAF is used in the European Copernicus program (CMEMS) for operational forecasting in the Baltic Sea. Also the German Maritime and Hydrographic Agency uses PDAF operationally (e.g. Bruening et al., 2021) and at the Chinese National Marine Environmental Forecasting Center (see e.g. Liang et al., 2019) applies it for sea ice forecasting. In this respect it is also unclear why 'operational NWP and related Earth system component DA' (line 66) for JEDI and OOPS is 'more general' (line 66) than 'applying ensemble DA algorithms to many different applications' (line 65). Actually, what should be 'more general' than a software that was developed for essentially any data assimilation application, as is the case for DART and PDAF? In comparison 'operational NWP and related Earth system component DA' appears to be more restricted. I can only recommend to avoid the impression that the authors intent to downgrade the value of systems like DART and PDAF, both of which are successfully used for real applications of 'Earth system component DA' and not just for 'applying ensemble DA algorithms to many different applications'. Apart from this, both DART and PDAF are obviously more 'mature' software compared to the very fresh development history of JEDI. Finally, PDAF is also not only providing 'ensemble DA

algorithms' (line 65), but also 3D-variational methods. Both DART and PDAF also provide tools for observation handling and diagnostics.

We thoroughly thank the reviewer for pointing out our mistakes regarding descriptions of other DA software that will be corrected in the revised version of the manuscript. The intent was certainly not to negatively judge any other DA software systems, but only to show how MIDAS is either similar or different from them in various ways. We have followed your suggestion and removed any specific claims about other systems to avoid making such errors. Thank you also for providing additional references.

Modified text replacing the previous description that contained errors: While, like MIDAS, these are all developed to cover a variety of DA applications, specific details about their range of functionality, software design, and current operational applications differ in many ways as compared with MIDAS. Like MIDAS, OOPS is currently used for operational NWP after being recently implemented at ECMWF (ECMWF, 2023).
[Modified text later concerning compilation of DA software and forecast model:] An important technical aspect of MIDAS is that its programs are executed separately from the forecast model software, while some of the other systems previously mentioned are compiled and executed together with the forecast model and exchange information between DA and forecast model through subroutines.

References:

Bruening, T., Li, X, Schwichtenberg, F., Lorkowski, I. (2021) An operational, assimilative model system for hydrodynamic and biogeochemical applications for German coastal waters. Hydrographische Nachrichten, 118, 6-15, doi:10.23784/HN118-01

Liang, X., Zhao, F., Li, C., Zhang, L., Li, B. (2020) Evaluation of ArcIOPS sea ice forecasting products during the nineth CHINARE-Arctic in summer 2018. Adv. Polar Science, 31, 14-25, doi:10.13679/j.advps.2019.0019

Response to Reviewer #2 (Reviewer's original text is in blue italics and author response is in black.)

Review of: The Modular and Integrated Data Assimilation System at Environment and Climate Change Canada (MIDAS v3.9.1), by Buehner et al (https://doi.org/10.5194/gmd-2024-55)

Reviewed by: C. Snyder, NCAR

Recommendation: Accept

This manuscript summarizes the design and implementation of a modular data assimilation (DA) system for Environment Canada, and gives example results. The writing is clear and concise and the topic is relevant for GMD. The manuscript shares refinements and approaches that will be of interest for other DA systems.

I offer comments for the authors' consideration, but I don't need to see the manuscript again.

We thank the reviewer for providing these useful comments that will be incorporated in the modified version of the manuscript.

1. *The MIDAS design embraces the simplicity that comes from divorcing the DA from model integrations and accepts the I/O overhead that comes with it. Leaving aside 4DVar, where there are separate reasons to include interfaces to the model integration, I have seen arguments from other efforts that the I/O overhead will be unacceptable, both for high-resolution ensemble DA and for fully coupled DA. It would be useful for the manuscript to include some perspective on this choice to rely on file I/O. Is it simply that you have built efficient parallel I/O?*

   For the current configurations of our deterministic and ensemble data assimilation systems, the relative time taken for I/O is not a major concern. This may, of course, change in the future due to increases in the model spatial resolution. The efficiency of I/O in MIDAS relies on the use of fully parallelized I/O and also ram disk to optimize the transfer of data between disk and memory. This will be better explained in the modified version of the manuscript. Also, examples of the times (both in terms of actual wall-clock time and percentage of the total execution time) will be provided.

   Text added to section 5: The entire process of reading and transposing the hourly 256-member ensemble takes approximately one minute of the total

execution time of about 11 minutes in the current global operational LETKF (described in section 6.1).

2. *Variable changes will be needed between model state, analysis variables, and variables required by observation operators How does MIDAS handle these? Could different models utilize the same code for variable changes?*

   MIDAS uses a set of variables for the main part of the DA algorithms that mostly correspond to the variables output and also read in by the forecast model. For some applications the background-error covariance matrix is specified in terms of different variables, including the log of specific humidity and stream-function/velocity potential. This variable transformation is handled directly within the corresponding B matrix module so that the rest of the MIDAS code does not need to be aware of these transformations. Any transformations between the variables used for the main part of the DA algorithms and the observed quantities are handled within the corresponding observation operator subroutines. If MIDAS were to be used with a different atmospheric model that produced a different set of variables, code modifications would be required to transform these variables, during the input and output stages, to the same set that are currently used.

3. *MIDAS has interesting capabilities within its observation operators, including the possibility of simulating based on slant paths and footprints that involve many model columns. I'd be interested to know more about how MIDAS handles the data distribution and parallelization in those cases. I don't see how the data distribution of Fig. 3 works with slant paths that cross multiple model layers, for instance.*

   The data distribution in Fig. 3 shows that each MPI task has an entire 2D field available in memory. This allows any type of horizontal interpolation to be applied, including the use of the observation footprint or slanted path. For slant path interpolation the vertical-level dependent latitude-longitude for all observations is used to perform the interpolation on each MPI task using the horizontal positions appropriate for the vertical level present on each MPI task. After the horizontal interpolation, the slanted column values for each vertical level are sent back to the MPI task where the rest of the observation operator will be performed (e.g. radiative transfer calculation or simple vertical interpolation). This is described in section 5, but with added details in the revised manuscript.

   Text added in section 5 (new text underlined): A different distribution is used when interpolating a 4D gridded state to observation location and time within the *stateToColumn_mod* module. For this, the gridded data are transposed into a distribution with respect to both variable type and vertical level, allowing the

interpolation to be performed on each MPI task for all time steps of a single complete horizontal field (Fig. 3). With the entire horizontal field available on a given MPI task, any interpolation approach and type of grid can be used without requiring additional MPI communication, including the use of a footprint operator to horizontally average many grid-points values or using level-dependent horizontal positions for interpolation to create a slanted column.

4. *MIDAS can be applied to a diverse set of applications. Unlike JEDI, DART, and PDAF, however, it is not (I think) designed to work interchangeably with different models in the same application. (E.g., swapping another atmospheric model for that used in the GDPS.) I'd be interested in discussion of that more limited scope and the MIDAS design. Does the limited scope permit simplifications or important design choices that are not possible in those other systems? Are those simplifications substantial?*

   Until now, the question of being able to use MIDAS with a different atmospheric model has not been considered. However, the recent modifications needed to enable ocean data assimilation with NEMO is somewhat analogous to the use of a different atmospheric model. The main modifications required were to add code for the reading and writing of the NEMO fields and also to add support for the tri-polar horizontal grid used by NEMO. The high-level design of MIDAS allowed these changes to be fairly straightforward to implement and a similar set of changes would likely be needed for supported a different atmospheric model. Therefore, I do not think that the use of a single atmospheric model led to any significant code simplifications.

5. *For JEDI citations, I suggest Liu et al. 2022 GMD as well, since it precedes Huang et al 2023. There is unfortunately no great reference for the underlying developments that support both those application papers; maybe Tremolet 2020 https://doi.org/10.25923/RB19-0Q26 ?*

   Thank you for the JEDI citation suggestions. I was not aware of the JCSDA publication dedicated to JEDI! These will be included in the revised manuscript.