Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-332-RC1, 2020

Geoscientific Model Development Discussions
Open Access
EGU

Interactive comment on "MLAir (v1.0) – a tool to enable fast and flexible machine learning on air data time series" *by* Lukas H. Leufen et al.

Anonymous Referee #1

Received and published: 23 November 2020

1 Overview

The manuscript describes a library which facilitates the development of end-to-end neural network workflows (though the name suggests it may support other ML algorithms) for time series forecasting (mostly focused on air quality).

Though the use case is fairly narrow in scope, the architecture of MLAir and the various features to validate the models using standard meteorological metrics is very interesting. MLAir's design also allows for development of reproducible ML pipelines which is essential if ML techniques are to be more widely used by the climate science community.

Printer-friendly version

Discussion paper

Overall, the aims and implementation of MLAir serve as a good template for future development of such frameworks and I consider it a valuable contribution to the geo-scientific community.

That being said, it seems to me that certain parts of the manuscript could be improved to enhance readability.

2 Major Comments

1. It would be helpful if the abstract clarified that MLAir is focused on neural networks and not any other kind of ML algorithm.

2. It would really help if the manuscript used different styles for conceptually different things. For instance, libraries (e.g., *TensorFlow*) are italicized, and so are class names (e.g., *Data Handler*). It would be preferable if the typewriter font is used to strengthen the correspondence with the figures (`DataHandler`). If *TensorFlow* is italicized, why is MLAir not italicized?

   If names like *Data Handler* refer to class names, then it is unclear why the names are split. If it describes the function performed by a piece of code, then it is not clear why it is italicized.

   Similar issues are found in the caption of Figure 1. Are *Experiment Setup, Preprocessing, Model Setup* class names or verbs? Is *Hyperparameter* a class name or terminology from ML? If it is the latter, why is it italicized, and why if it is the former, then why is it not mentioned in Line 91? Similarly, the text used in the various bubbles in Figure 1 itself could benefit from such formatting (is *Run Environment* a class name or a simply a description?).

   As it stands, the manuscript's treatment of different terminologies is too confusing for the reader to precisely understand what each term stands for. The manuscript

would also benefit from a sentence or two stating the typographical choices made by the authors.

3. Section 2.1 would benefit from a definition of the main components – *Run Environment, Data Handler, Model Class* – what they do, and how they fit into the design of MLAir. Currently, only *Run Environment* is described, and *Data Handler* has a fairly confusing description (Line 93: Data Handler is responsible for an accurate use of the data).

   I strongly recommend rewriting Section 2.1 with the reader in mind, to ensure that readers are systematically introduced to the concepts and/or stages underlying MLAir and the corresponding software components that implement these concepts and/or stages.

4. To ensure the a reader with no background in ML is able to relate to the design of MLAir, a brief introduction to a typical ML workflow (test, train, validation, epoch, hyperparameter, etc.,) would be really helpful in Section 2.1.

5. In Section 2.2, *window_history_size* is not defined anywhere. Therefore, what it does and how it might impact network architecture (Line 128–129) is not clear.

6. Line 142: The term epoch is used without defining it anywhere. Definitions and motivations for such terms could be added to the brief description of a typical ML workflow that was mentioned previously.

7. Line 220: If the skill score is defined as a ratio of a metric like MSE, how does one obtain positive and negative skill?

8. Line 229: The process of bootstrapped predictions could be explained a little more in detail for completeness. “the time series of each individual input variable is resampled n times (with replacement) and then fed to the trained network.” was not sufficient for me (at least) to understand what was being attempted and achieved.

9. Figure 12: I might have missed something, but it is unclear what AI, BI, CI, CASEI etc., are. Furthermore, it is unclear what “terms” and “contributing terms” (Line 226) are.

10. Lines 417-421: The line numbers in Figure 14 that are referred to here don’t seem to correspond to their respective descriptions. I hope the authors verify the same for other such figures as well.

11. Line 440: A more detailed description of how the imbalance problem is addressed would benefit the reader who wishes to use this feature.

3 Minor Comments

1. Line 58: It also allows to deploy → It also allows deploying

2. Line 59: use of GPUs . I think it is proper to acknowledge that usage of GPUs is due to the underlying tensorflow library.

3. Line 60: Concurrent to a simple usage with low barriers for ML-callow scientists,: **Not sure what is meant by this sentence**.

4. Line 84: as many customization → To facilitate customization(?)

5. Line 83: Why is Workflow capitalized?

6. Figure 1: exemplary: **exemplary usually means commendable. I don’t think that is what the authors meant.**

7. Figure 6: exemplary measurement stations: **Same as above. This applies to all other places where exemplary has been used.**

8. Line 110: daily aggregated: **does this mean daily mean? daily max? It is unclear how the aggregation is done.**

9. Line 168: Beside the file, which contain the model → Besides the file, which contains the model

10. Line 181: intended to add own graphics in MLAir → intended to add custom graphics in MLAir

11. Line 185: major waters → major water bodies

12. Line 191: are meant to give an insight into → are meant to provide insight into

13. Line 195: each month separately as box-and-whisker → each month separately as a box-and-whisker diagram

14. Line 205: marginal distribution is shown as histogram (lite grey) → marginal distribution is shown as a histogram (light grey)

15. Line 242: independent on the OS → independent of the OS

16. Line 350: Since the spatial dependency of two distinct stations may variegate related → Since the spatial dependency of two distinct stations may vary(?) related

17. Line 433: how to implement a ML model → how to implement an ML model

18. Line 436: is to assume an independent and identically distribution and therefore augment and randomly shuffle data to produce a larger number of input samples with a broader variety. → is to assume independent and identically distributed data and therefore augment and randomly shuffle the data to produce a larger number of input samples with a broader variety. **Consider rewriting this sentence to enhance readability.**

19. Line 442: To address this issue, MLAir allows to place more → To address this issue, MLAir allows placing more

---

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-332, 2020.

Printer-friendly version

Discussion paper