ENS-10 Datasheet

In this section, we follow the *Datasheets for Datasets*[6] instructions to provide documentation for our dataset. To this end, in separate sections, we answer the questions for Motivation, Composition, Collection Process, Preprocessing/Cleaning/Labeling, Uses, Distribution, and Maintenance of the dataset. In each section, we use **bold** and *italic* fonts for the main and following questions.

6 Motivation

- **For what purpose was the dataset created?** *Was there a specific task in mind? Was there a specific gap that needed to be filled?*

  The ENS-10 dataset is created to help the machine learning community develop more sophisticated data-driven solutions for ensemble post-processing of weather prediction models.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

  The dataset is created in collaboration between the European Centre for Medium-Range Weather Forecasts (ECMWF[7]) and the Scalable Parallel Computing Lab (SPCL[8]) at ETH Zurich. The data are extracted by Peter Dueben.

- **Who funded the creation of the dataset?** *If there is an associated grant, please provide the name of the grantor and the grant name and number.*

  This dataset was funded by the MAELSTROM EuroHPC project. The MAELSTROM project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955513. The JU receives support from the European Union's Horizon 2020 research and innovation programm and United Kingdom, Germany, Italy, Luxembourg, Switzerland, Norway.

- **Any other comments?**

  No.

7 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions be- tween them; nodes and edges)?*

  ENS-10 consists of different variables representing the atmospheric state in a fixed-point numerical representation. Each data sample is associated with a date, with three lead times T=0, 24, and 48 hours. For each lead time, the data has all variables in the dataset at each pressure level and grid point. The variables are saved in GRIB format.

- **How many instances are there in total (of each type, if appropriate)?**

  The dataset consists of 105 ensemble forecasts for each of the years 1998–2017. Each ensemble forecast consists of ten ensemble members, the initial condition (lead time T=$0\,\mathrm{h}$) for each ensemble member, and two forecasts at lead times T=$24\,\mathrm{h}$ and T=$48\,\mathrm{h}$. The dataset has 2100 instances in total.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

  The dataset is a contiguous-time subset of 20 years. Spatially, it encompasses the entire Earth and is discretized onto a structured grid with $0.5°$ resolution and 12 pressure levels.

[6] https://arxiv.org/abs/1803.09010
[7] https://www.ecmwf.int
[8] https://spcl.inf.ethz.ch

With respect to the atmospheric state, it is a representative subset of 18 out of all the internal fields in the original numerical model.

- **What data does each instance consist of?** *"Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.*

  Each instance consists of the raw data in GRIB format, including the parameters in fixed-point representation (as is output by the weather model), over all the spatial points and ensemble members.

- **Is there a label or target associated with each instance?** *If so, please provide a description.*

  The target of the ENS-10 prediction correction task is defined by matching the output to a second ground-truth dataset (ERA5). The goal of the task is to minimize uncertainty in prediction over the two nonzero lead times (24, 48 hours) over the ensemble member distribution.

- **Is any information missing from individual instances?** *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

  We do not include all the possible fields (variables) in our dataset. We only provide a representative of relevant fields to represent the atmospheric state for the prediction correction task, due to the large number of fields and consequent storage size.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** *If so, please describe how these relationships are made explicit.*

  No.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** *If so, please provide a description of these splits, explaining the rationale behind them.*

  Yes. To mimic the ideal learned system, the recommended split is to take the last two years (2016–2017) as the test set, training over historical data (1998–2015). This ensures that the system accounts for, e.g., shifting climate over time. Among the training set, one or two arbitrary years can be chosen as a validation set, enabling cross-validation.

- **Are there any errors, sources of noise, or redundancies in the dataset?** *If so, please provide a description.*

  There are a number of sources of uncertainty in numerical weather prediction, either from the underlying data or choices within the model. Key sources include data (or aleatoric) uncertainty, from noise in observational measurements; and model uncertainty, stemming from structural (e.g., choice of forecast model) and parametric uncertainties. Our benchmark experiments further assume a Gaussian distribution for prediction correction, which need not hold in practice, although other users of ENS-10 need not hold to this assumption. We further note that the goal of the prediction correction task we use ENS-10 for is to correct the output of a NWP model, not to directly operate on meteorological measurements.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

  The dataset is self-contained. To establish ground-truth for the prediction correction task, we use the ERA5 dataset.

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?** *If so, please provide a description.*

  No.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** *If so, please describe why.*

  No.

8 Collection Process

- **How was the data associated with each instance acquired?** *Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

  The data is the output of ensemble simulations with the Integrated Forecast System (IFS), a global weather prediction system. The data was produced using the production weather prediction workflow of ECMWF, collecting millions of weather observations across the globe from various different sources, including satellites, weather balloons, ships, planes, and ground measurements. The observations are then assimilated into the forecast model to generate initial conditions, and the forecast model is used to predict the weather of the future using supercomputing facilities. Finally, the model data is mapped onto the grid used for this paper. The truth data that is used for training is so-called re-analysis data which is based on a data assimilation approach to derive the most consistent state of the atmosphere combining information of observations and model simulations for a specific point in time.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** *How were these mechanisms or procedures validated?*

  The primary mechanism to produce the data is the use of the IFS numerical weather prediction system to perform forecast simulations.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

  The dataset is a sample of a larger dataset in the sense that not all timesteps or physical variables of the IFS model simulations have been extracted. This is to keep the dataset to a reasonable size (several terabytes).

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

  The data are collected by Peter Dueben. The model that was used to generate the data was developed by a very large group of scientists, mainly based at ECMWF [9].

- **Over what timeframe was the data collected?** *Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time- frame in which the data associated with the instances was created.*

  Simulations have been performed for a timeframe of 20 years of weather situations. However, the data assimilation and model simulations were running operationally at ECMWF and extracted within a short period of time.

- **Were any ethical review processes conducted (e.g., by an institutional review board)?** *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

  No.

9 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** *If so, please provide a description. If not, you may skip the remaining questions in this section.*

  The model output data is stored in GRIB files using the standard GRIB number format (fixed-point scaled integers) for all output fields. This is used for all simulations of the atmosphere at ECMWF and is not known to reduce the information content. The model output data was mapped from the unstructured grid used by the simulation onto a structured longitude/latitude grid using ECMWF's standard mapping procedure for meteorological fields.

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *If so, please provide a link or other access point to the "raw" data.*

  No.

- **Is the software that was used to preprocess/clean/label the data available?** *If so, please provide a link or other access point.*

  The Meteorological Archival and Retrieval System (MARS) was used to retrieve the model data from ECMWF's data archive and to perform the mapping to the structured longitude/latitude grid[9].

- **Any other comments?**

  No.

10 Uses

- **Has the dataset been used for any tasks already?** *If so, please provide a description.*

  The dataset was used for the *Prediction Correction* task. See Section 3 and Grönquist et al. [15].

- **Is there a repository that links to any or all papers or systems that use the dataset?** *If so, please provide a link or other access point.*

  Yes. We provide such information on our GitHub repository[10].

- **What (other) tasks could the dataset be used for?**

  The dataset could be used for correcting the prediction of any environmental variables.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** *For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

  No.

- **Are there tasks for which the dataset should not be used?** *If so, please provide a description.*

  We discuss some limitations in Section 5; in particular, the dataset is not suitable for weather *prediction* on its own.

- **Any other comments?**

  No.

11 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** *If so, please provide a description.*

  Yes. The dataset is freely available as discussed below.

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?***Does the dataset have a digital object identifier (DOI)?*

  The dataset is accessible directly from
  https://storage.ecmwf.europeanweather.cloud/MAELSTROM_AP4/
  or using a Python APIs from https://github.com/spcl/climetlab-maelstrom-ens10.

- **When will the dataset be distributed?**

  The dataset is available to download as of the submission deadline. All usage examples will be ready by June 30, 2022 at https://github.com/spcl/ens10.

[9] https://confluence.ecmwf.int/display/UDOC/MARS+user+documentation
[10] https://github.com/spcl/ens10

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

  The ENS-10 dataset is available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. The license description can be accessed at https://storage.ecmwf.europeanweather.cloud/MAELSTROM_AP4/LICENCE.txt.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

  No.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

  No.

- **Any other comments?**

  No.

12 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**

  The dataset is hosted on the ECMWF servers and mirrored at ETH Zürich. The dataset will be maintained by ECMWF and by the Scalable Parallel Computing Laboratory (SPCL) at ETH Zürich.

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

  The dataset maintainers can be contacted via issues or pull requests on the ENS-10 Github repository.

- **Is there an erratum?** *If so, please provide a link or other access point.*

  No.

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** *If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?*

  The dataset is complete and will not be updated. The production version of the Integrated Forecast System, which produces the hindcasts from which ENS-10 is derived, changes over time. Hindcasts are also only produced for a fixed number of years back until the present. Thus, the meteorological fields produced would change, and newly generated data would not be suitable for use in prediction correction.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** *If so, please describe these limits and explain how they will be enforced.*

  No.

- **Will older versions of the dataset continue to be supported/hosted/maintained?** *If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

  There is no older version of the dataset.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

  Any modification and extension of the dataset under the Creative Commons license is permitted.

- **Any other comments?**
  No.

13 Responsibility

The authors bear all responsibility for violations of rights related to the ENS-10 dataset.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We discuss this in Section 5. In particular, ENS-10 is not suitable for weather prediction, and inherits any biases present in the underlying numerical weather prediction models that generated its data.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] While any dataset could potentially have negative social impacts, we do not anticipate any harms from introducing a dataset for post-processing ensemble weather forecasts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The dataset we propose, ENS-10, is available through links in Section 5 and is further discussed in the appendices. Code and instructions for our baselines is included in the supplementary material and our dataset repository.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] These are discussed in Section 4. Full details to reproduce the results are in our supplementary material and dataset repository.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 4.2. We repeat each experiment three times and report their mean and standard deviation.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.2.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] The construction of the ENS-10 dataset is described in Section 2. Section 4.1 describes our baseline models, which are standard methods we implemented and adapted from prior work.
   (b) Did you mention the license of the assets? [Yes] ENS-10 is licensed under a CC BY 4.0 license.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include links for the ENS-10 dataset and associated baselines in Section 5 and provide additional details in the appendix.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] The ENS-10 dataset does not contain data from people; it is produced from numerical weather prediction models.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] ENS-10 does not contain any personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]