REQUEST FOR A SPECIAL PROJECT 2024–2026

MEMBER STATE: Denmark

Principal Investigator[1]: Martin Stendel, Senior Scientist, Ph.D.

Affiliation: Danish Meteorological Institute

Address:
Lyngbyvej 100

DK 2100 Copenhagen

Other researchers: Esben Haubro Skov, Scientist, National Archive of Denmark and Danish Meteorological Institute

Project Title: ROPEWALK – Rescuing Old data with People's Efforts: Weather and climate Archives from LogbooK records

……………………………………………………………………………………………………….

To make changes to an existing project please submit an amended version of the original form.)

If this is a continuation of an existing project, please state the computer project account assigned previously.	**SP** ………….	
Starting year: (A project can have a duration of up to 3 years, agreed at the beginning of the project.)	2024	
Would you accept support for 1 year only, if necessary?	YES ☒	NO □

Computer resources required for project year:		**2024**	**2025**	**2026**
High Performance Computing Facility	[SBU]	500000	500000	500000
Accumulated data storage (total archive volume)[2]	[GB]	1000	1000	1000

EWC resources required for project year:		**2024**	**2025**	**2026**
Number of vCPUs	[#]	128	128	128
Total memory	[GB]	256	256	256
Storage	[GB]	1000	1000	1000
Number of vGPUs[3]	[#]	8	8	8

Continue overleaf.

[1] The Principal Investigator will act as contact person for this Special Project and, in particular, will be asked to register the project, provide annual progress reports of the project's activities, etc.

[2] These figures refer to data archived in ECFS and MARS. If e.g. you archive x GB in year one and y GB in year two and don't delete anything you need to request x + y GB for the second project year etc.

[3] The number of vGPU is referred to the equivalent number of virtualized vGPUs with 8GB memory.

Principal Investigator: Martin Stendel, Senior Scientist, Ph.D.

Project Title: ROPEWALK – Rescuing Old data with People's Efforts: Weather and climate Archives from LogbooK records

Extended abstract

All Special Project requests should provide an abstract/project description including a scientific plan, a justification of the computer resources requested and the technical characteristics of the code to be used. The completed form should be submitted/uploaded at https://www.ecmwf.int/en/research/special-projects/special-project-application/special-project-request-submission.

Following submission by the relevant Member State the Special Project requests will be published on the ECMWF website and evaluated by ECMWF and its Scientific Advisory Committee. The requests are evaluated based on their scientific and technical quality, and the justification of the resources requested. Previous Special Project reports and the use of ECMWF software and data infrastructure will also be considered in the evaluation process.

Requests exceeding 5,000,000 SBU should be more detailed (3-5 pages).

The proposed special project aims at digitizing all weather observations in ship journals and logbooks, which are stored in Rigsarkivet, the National Archive of Denmark. A huge amount of data (almost one shelf kilometer) is stored, beginning as early as 1675. With the exception of the Napoleonic wars and the Danish state bankruptcy in 1814, the data is complete. In particular, there were no losses during the Second World War.

The collection in the archive is remarkable for several reasons. In the archive, logbooks from Danish ships over large parts of the Northern Hemisphere are stored. Of particular interest are observations from two regions, the Øresund and Greenland.

Øresund Dues

In connection with the Sound Dues which every ship passing the sound or belts had to pay between 1426 and 1857 (Fig. 1), weather observations were made on board of war ships placed at strategic locations near Copenhagen, Helsingør and Nyborg. These ships had to ensure that no one passed without paying. The economic importance of the dues was enormous and made up almost half of the Danish state income in the 17th century. Up to 1703, the money was directly handed over to the king.

For practical reasons, weather observations were tabulated as early as the first half of the 18th century. In several cases, observations were conducted every time the ship bell was struck, resulting in 48 observations in the course of one day. The early part of the logbook collection is from the Little Ice Age, and numerous ice observations in the Danish waters have been preserved.

Greenland Voyages

Like other seafaring nations, Danish merchant ships made voyages to their colonies, in this case (western) Greenland. The Greenlandic Trade Company had a monopoly for commerce with Greenland for nearly 200 years, which was enforced strictly, such that foreign ships would not be allowed to call a Greenlandic port. These "Greenland Voyages" were conducted up to several dozen times per year. In many cases, detailed sea ice observations have been made.

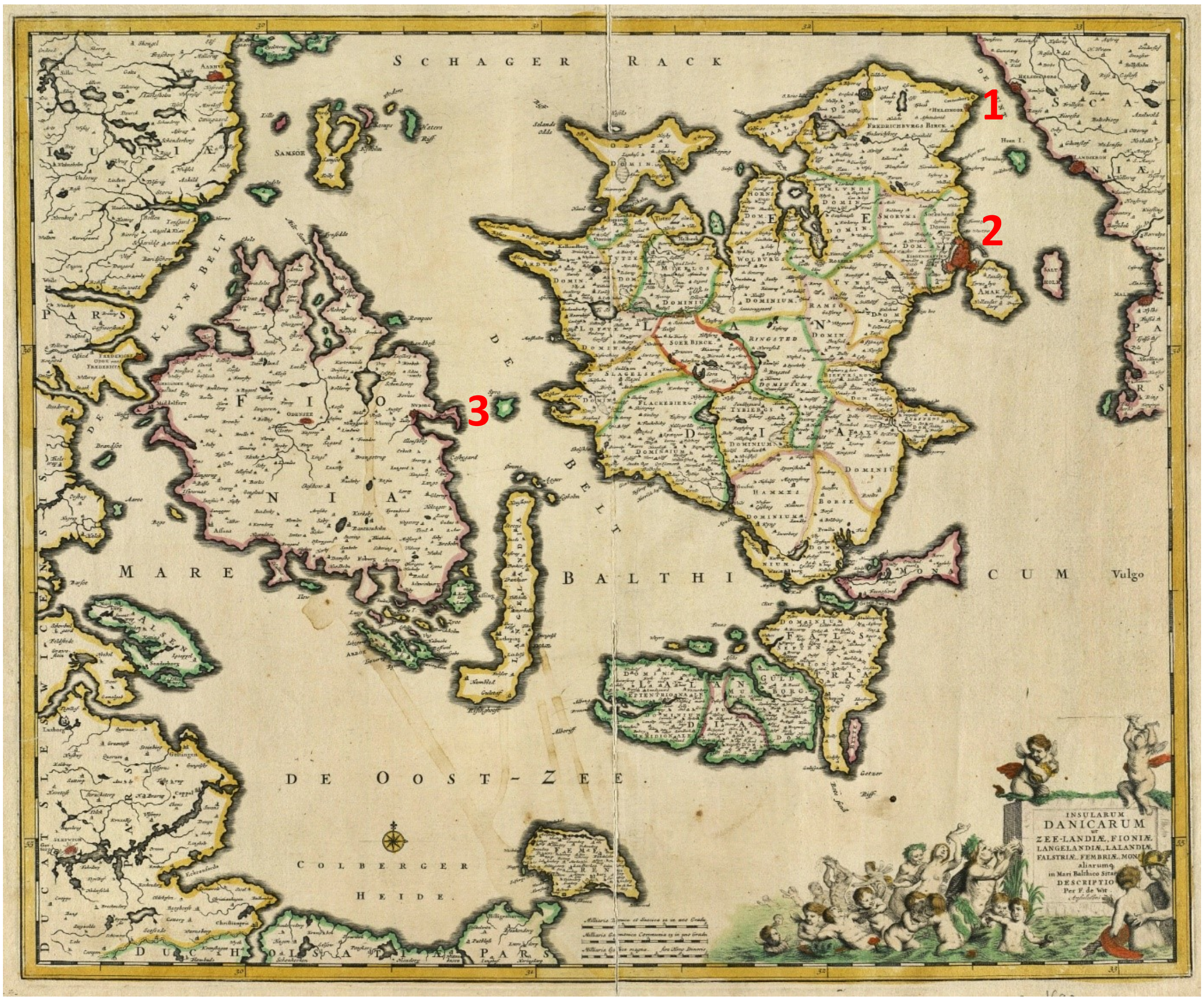
SCHAGER RACK
1
2
3
FIONIA
ZEELANDIA
MARE BALTHICUM Vulgo
DE OOST-ZEE
COLBERGER HEIDE
INSULARUM DANICARUM ZEE-LANDIÆ, FIONIÆ, LANGELANDIÆ, LALANDIÆ, FALSTRIÆ, FEMBRIÆ, MONÆ aliarumq. in Mari Balthico Sitar DESCRIPTIO Per F. de Wit

Figure 1: Frederick de Witt: Insularum Danicarum ut Zeelandiæ, Fioniæ, Langelandiæ, Lalandiæ, Falstriæ, Fembriæ, Monæ aliarumq. in Mari Balthico sitar (Map of the Danish islands, namely Sealand, Funen, Langeland, Lolland, Faster, Fehmarn, Møn and others); Amsterdam 1670. Royal Library. Locations of war ships on duty are marked off Helsingør (1), Copenhagen (2) and Nyborg (3).

Transcription

The Danish Meteorological Institute (DMI) and the National Archive of Denmark are currently involved in a project (ROPEWALK; 2023-2026) to digitize and transcribe all weather observations in ship journals and logbooks stored in the archive.

The original ship journals and logbooks are currently being scanned by the scanning department of the National Archive in very high resolution. Figure 2 shows two examples. The scans will then be transcribed by means of machine learning. This is possible, not least since the political system in Denmark was absolutistic between 1660 and 1848, and logbooks from different periods resemble each other much more than is the case for the nautical heritage in other seafaring nations. Where machine-enhanced transcription is not possible, the data will be transcribed with the help of volunteers.

All transcribed data will be made publicly available. They will be used as input for reanalysis projects or for other future research.

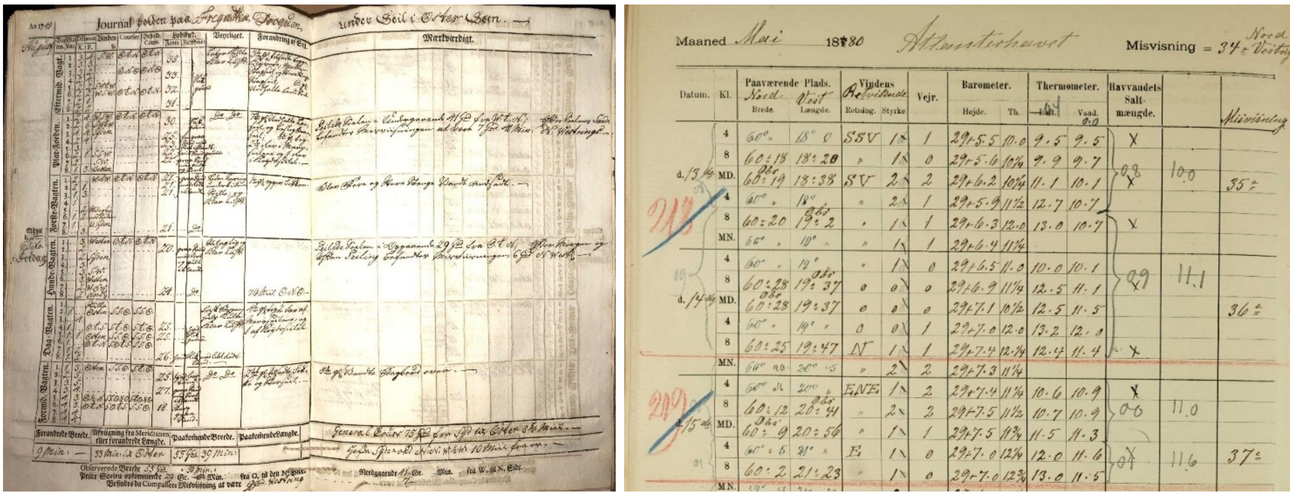

Figure 2: Examples of ship journals to be transcribed. Left: Frigate Docquen, sailing in the Baltic Sea in August 1748. Note the preprinted tables. Right: Brig Hvalfisken, sailing on the Atlantic in May 1880. Pressure and SST observations are readily available.

Requested HPC resources

The objective of the project is to make handwritten data observations applicable to the climate modelling community by extracting information from images. The data extraction process involves computationally intensive tasks, and optimal computation specifications are required to achieve accurate results.

To accomplish the data extraction, we utilize computer vision and machine learning techniques for Handwritten Text Recognition and Semantic Segmentation, specifically Convolutional Neural Networks (CNNs) and Transformers models such as RESNET, U-net, Segformer, TrOCR, Tesseract, among others. The training process for these models, particularly the largest ones, demands a significant amount of GPU memory, estimated to occupy approximately 48-64 GB of GPU space. We prefer to utilize a single GPU node or alternatively a setup that supports parallel GPU computing for this purpose.

Furthermore, multiple CPU cores are required for inference and general image processing tasks. Therefore, we request a setup with multiple CPU cores to ensure efficient processing of the data. As for storage, while our full-scale image project dataset contains multiple terabytes (TB) of data, we estimate that 1 TB of disk space would be sufficient to store and manage the necessary data. For code execution, we will use Python and common libraries e.g. Tensorflow, PyTorch (lightning) and OpenCV.

We have based our requested technical specifications on the information provided on the webpage (ECMWF 2023) and have scaled up the configurations to achieve a 64GB GPU. A GPU-enable instance has the following specification: 16 GB GPU, 16 CPU and 32 GB memory.

In light of the above, we kindly request the following technical specifications for our compute resources:

- GPU Space: 64 GB
  - Either one powerful GPU or multiple GPUs that support parallel computing
  - CPU: 128 vCPU[2] - 16 CPUs multiplied by 8 gives us a total of 128 vCPUs.

[2] Calculation based on https://confluence.ecmwf.int/display/EWCLOUDKB/GPU+support+at+ECMWF and scaled up to 64 GB GPU. We are aware that this is above the "Maximum per Special Project per year" limit.

  - Based on the estimate of resources, we therefore ask for access to the GPUs as well.

- However, we have identified an existing computer resource that might fulfil our needs with the following specifications: Nvidia A6000 GPU (48 GB), Intel Xeon 6242R CPU (20 cores, 40 threads), 128 GB memory. Therefore, less cores may be sufficient. It is unclear to us if a comparable solution is available at ECMWF. Please confirm if our calculation for vCPUs is correct or if a lower number of vCPUs would suffice.
  - Memory: 256 GB (32 GB multiplied by 8 gives us a total of 256 GB of memory).
- Disk Space: 1 TB
- Program language: Python
- Programming package: Tensorflow, PyTorch Lightning and OpenCV

Remark on SBUs: To our best knowledge, a similar approach has not yet been conducted on ECMWF's HPC. This means we unfortunately do not have an estimate about how many SBUs per year are needed. So it might be necessary to adjust the number of billing units in the course of the project.

We therefore expect the required number of SBUs per year on the order of 500000.

Scientific outcomes

The ship journal and logbook data are unique under several aspects.

- The Royal Greenlandic Merchant Company had a monopoly on trade with Greenland for almost 200 years. That means that almost all existing observations in the 18th and the first half of the 19th century are stored in the National Archive.
- Millions of observations related to the Øresund duties exist. Their temporal and spatial resolution is so high that it appears possible to create a regional reanalysis for the 18th and 19th centuries.

More generally, the transcribed data will serve as input to reanalyses. They will thus serve the entire climate modelling community.

References

GPU support at ECMWF. (2023, 06 22). Retrieved from https://confluence.ecmwf.int/display/EWCLOUDKB/GPU+support+at+ECMWF