Archives and climate science: transforming paper documents into global climate datasets

Wilkinson, C. ORCID: https://orcid.org/0009-0009-6004-6995, Divine, D. V. ORCID: https://orcid.org/0000-0003-0548-6698, Brohan, P., Brönnimann, S., Compo, G., Hawkins, E. ORCID: https://orcid.org/0000-0001-9477-3677 and Slivinski, L. C. ORCID: https://orcid.org/0000-0002-3531-3889 (2025) Archives and climate science: transforming paper documents into global climate datasets. Norsk arkivforum, 31 (1). pp. 9-17. ISSN 2387-2829 doi: 10.18261/naf.31.1.3 Available at https://centaur.reading.ac.uk/122966/

Identification Number/DOI: 10.18261/naf.31.1.3
<https://doi.org/10.18261/naf.31.1.3>

Publisher: Scandinavian University Press

Norsk Arkivforum | Universitetsforlaget

Volum 31, nr. 1-2025, s. 9–17
ISSN online: 2387-2829
FAGARTIKKEL
DOI: https://doi.org/10.18261/naf.31.1.3

Archives and Climate Science: Transforming Paper Documents into Global Climate Datasets

Clive Wilkinson
Research Associate, University of Reading, UK
clivewm.wilkinson@gmail.com
https://orcid.org/0009-0009-6004-6995

Dmitry V. Divine
Senior Researcher at the Norwegian Polar Institute, Tromsø, Norway
Dmitry. Divine@npolar.com
https://orcid.org/0000-0003-0548-6698

Philip Brohan
Met Office, Exeter, UK
philip.brohan@metoffice.gov.uk

Stefan Brönnimann
Prof. Dr., University of Bern, Switzerland

Gilbert Compo
Senior Research Scientist, NOAA Physical Sciences Laboratory Modeling & Data Assimilation Division
gilbert.p. compo@noaa.gov

Ed Hawkins
Professor, National Centre for Atmospheric Science, University of Reading
e.hawkins@reading.ac.uk.
https://orcid.org/0000-0001-9477-3677

Laura C. Slivinski
Research Scientist, NOAA/OAR/PSL, Boulder CO, USA
laura.slivinski@noaa.gov
https://orcid.org/0000-0002-3531-3889

Abstract

Climate science depends on reliable historical weather data. Much of this data can be found in paper documents spanning decades or centuries and is held in archives around the world. Norway has a rich archival heritage, of which only a small part has so far been used by climate scientists. The data is needed to establish long-term climate trends, for the reconstruction of past weather events, to better understand the best sites to establish wind and solar farms and to improve weather forecasting models as well as many other applications. Documents from the archives are scanned and transcribed by various methods with the data then subjected to rigorous quality control, before being used for global climate datasets and weather reconstructions. Finding and assessing suitable material, then imaging and transcribing the documents, is labour-intensive work, and the scale of the task is daunting if the scientific community is to fully realise the potential of archives in Norway and elsewhere in the world. This provides a unique opportunity for archives and archivists to work with scientists from around the world to improve

our understanding of the earth's climate and thereby mitigate some of the effects of climate change and improve weather forecasting.

Keywords
Climate, weather, document transcription, data rescue, whaling, artificial intelligence

The year is 1937. Under a leaden sky, a large Norwegian ship rises slowly and then falls ponderously onto the trough of a long heavy swell. A cold wind blows flurries of snow onto the deck, while the sea surface is covered for miles around with small lumps of ice. Far to the south, beyond the ice barrier, lies the continent of Antarctica. Alongside, a whale catcher, dwarfed by a large factory ship, bumps gently alongside in the swell, a whale in tow. In his tiny office, the ship's writer or clerk hears the commotion on deck, heralding the receipt of yet another whale for processing and the additional paperwork it entails. This is the hidden side of life at sea; the endless paperwork needed to fulfil maritime regulations, and the requirements of an increasingly regulated whaling industry.

As well as the completion of the ship's navigational logbook – a requirement for all vessels – the ship's owner A/S Thor Dahl, from Sandefjord, required a mass of documentation, much of it indicating compliance with the international whaling regulations that Norway and other nations had recently agreed to observe. Each and every whale capture had a separate record, detailing the species, the position of capture, the weather, sea and ice conditions. This data then had to be transferred to a whale catch book or *fangstdagbok*, which also kept a daily record of weather and ice conditions. There was a host of other documentation, and if this was not already enough, someone had volunteered this particular overworked ship's clerk to fill in forms for the United States Hydrographic Department, recording ocean currents, sea temperatures and ice conditions.

Every whaling ship, whether Norwegian, British, American, South African or German or some other nation, was required to submit the same documentation. This documentation was preserved in the shipping company archives, and ultimately the state and regional archives in the host nations of the whaling companies. These specialised collections preserving the history of whaling are just a small part of the maritime collections found across the globe and held in state and regional archives, museums, libraries, universities and scientific institutions. The ship's clerk could never have imagined that the information he helped to collate would one day be of critical value to meteorologists, oceanographers and climate scientists around the globe.

Historians will tell you that in order to understand the present and to adequately plan for the future, you must first understand the past. The same applies to climate science. Without data from the past, it would be impossible to assess present climate trends and extremes.

Just how critical this data is, can be demonstrated by one astonishing fact. If it were not for the pelagic fishing of Antarctic waters in the 1920s and 1930s, mostly by Norwegian and British flagged whalers, modern science would have no record whatsoever of the weather and ice conditions of the far south at that time. In fact, prior to the 1950s there are significant gaps in the global climate record for many parts of the world, a situation that worsens the further back in time one goes. However, the data do often exist, in original paper form in archives and museums around the world. Some collections of marine and terrestrial data are known and have been partially exploited, but many potentially important collections of weather data remain unknown to science. Recent research has uncovered some of these collections, but the scientific community needs the help of archivists to find more data.

So, how is this historical data used by scientists, how do we convert paper documents into a digital form for scientific analysis, and importantly, what types of documents do we

need to find and process in order to improve the observational record, both geographically and over time?

How Do Scientists Use the Historical Observations of Weather?

By collecting weather data from archives around the world it is possible to collate the information into datasets and atmospheric reanalyses (or computer-generated maps) of past weather. It is from processing and analysing these historical records that scientists now understand that the planet is undergoing a period of global warming.

Why Do We Need Historical Records of Air and Sea Temperatures?

Our knowledge of how the climate has already changed, and the role of human influences on the climate, relies on observations. More than a billion individual thermometer measurements make up our current archives of digitised temperature observations dating back to, in some cases, the late 1700s. However, there are gaps in our knowledge and it is likely that at least the same number of weather observations remain unavailable to science in various paper archives around the world.

For 1937 – the year described in the introduction – there are currently very few sea surface temperature observations from the southern Pacific (see Figure 1) and none in the Southern Ocean or Arctic. The main shipping trade routes are clearly visible. Efforts are ongoing to find, image, digitise, quality-control and integrate additional observations from many sources, such as whaling ship archives, into global databases in order to fill some of these gaps. The picture can be made more complete, but some gaps can never be filled as no ship went there at that time.

On land, it is a similar story. Many observations are available, especially in Western Europe and other developed parts of the world. The largest gaps remain in Africa, South America and parts of Asia.

What Are Atmospheric Reanalyses and How Are They Used?

Atmospheric reanalyses are datasets of four-dimensional reconstructions of historical weather. Here, the fourth dimension is time. These datasets are essentially a huge collection of digital weather maps available at regular time intervals (usually 4–8 times per day) for a period spanning decades into the past. They provide maps of temperature, winds, humidity, precipitation, cloudiness, and other weather-related variables, not just at the surface of the earth but high up into the atmosphere. These datasets are created by combining historical weather observations with a priori physical knowledge of the earth system obtained from weather forecast models. Some reanalyses, like the NOAA-CIRES 20th Century Reanalysis (Compo et al., 2011; Slivinski et al., 2019), use particular methods to produce these maps as far back as the early nineteenth century; currently, the lack of available historical observations limits going back further.

Reanalyses are distinct from raw observations. While useful, observations on their own cannot provide a consistent record of weather and climate without gaps in both space and time. Instruments change, their locations may move, and sometimes the surrounding environment changes significantly, such that painstaking and imperfect fixes are needed to make these records homogeneous. Importantly, observations do not exist everywhere, continuously in time. Reanalyses not only fill these gaps but do so consistently in space and time, which allows for more reliable studies of climate and trends. For example, reanalyses can be used for renewable energy planning (Wilczak et al., 2024): the historical intensity and frequency of wind or solar droughts in a particular unobserved region can inform a

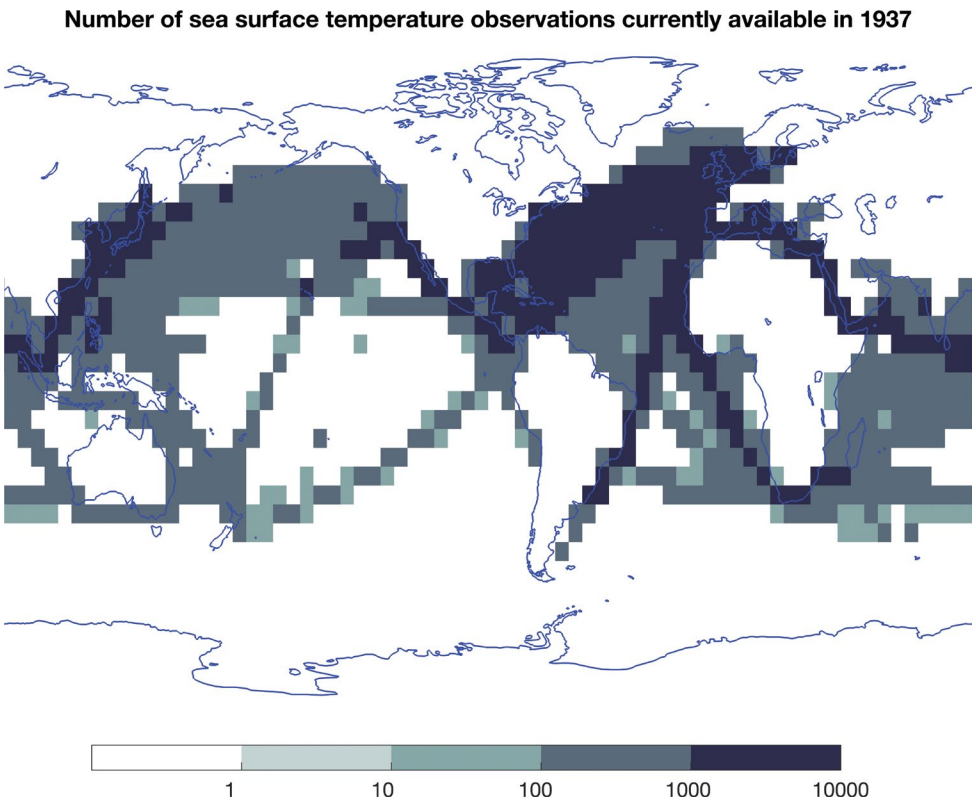

Figure 1. The need for more historical air and sea temperature observations is essential, but there is another parameter that in recent decades has become just as important. These observations are atmospheric pressure, from a variety of measuring instruments such as mercury and spirit barometers, aneroid barometers and other devices. Historical pressure observations can be put into a weather forecast model, along with other parameters such as air and sea temperatures, wind velocity and direction, and other variables, to produce what is termed an atmospheric reanalysis or retrospective analysis of past weather.

cost-benefit analysis of building a wind or solar farm there. In addition, reanalyses can be used to investigate how the intensity and frequency of impactful events like hurricanes, floods, droughts, blizzards, heatwaves, and wind storms are changing in time, which can allow communities to better prepare for future extreme events (e.g., Burn & Palmer, 2015; Chand et al., 2022; Donat et al., 2016; Hawkins et al., 2023). Finally, reanalyses have recently been the key to training weather forecast models produced by machine learning (ML; e.g., Bi et al., 2023; Lam et al., 2023; Pathak et al., 2022). These ML models have shown an impressive ability to produce accurate weather forecasts in a fraction of the time needed for traditional physics-based weather models.

From Archive to Digital Dataset: Exploring the Archives, Scanning and Imaging

Turning paper records into usable climate data is a process that first requires a person to find suitable material in the archives. These persons are 'data rescue' experts who must deploy the skills of the historian, the archivist and the scientist to explore an archive, then find and critically assess the material. Selected material must then be scanned or photographed, with the resulting images organised and catalogued and prepared for transcription. This process requires close cooperation with the archives and with archivists. Most scientists have neither the time nor the expertise to undertake this specialised work, meaning that archivists and data-rescue experts are critical to the success of this endeavour.

Document Transcription: Crowdsourcing and Artificial Intelligence

Climate records, for example ship logbook pages, are appealing and powerful: densely packed with information, containing precise quantitative records of temperature, pressure, position and time, as well as less structured comments. To a human eye, all this information is well laid out and readily accessible, but for climate scientists, pages, and page images, are not what we need.

Climate reconstructions depend on computer-readable databases of observations: tables of numbers stored in indexed and searchable online archives. So, to use the ship logbooks and other weather records, we have to transcribe the pages – to read each page, and enter all the information on the page into a database. It does not sound very difficult, but the transcription process is a major barrier to the scientific use of archival records.

Given an appropriate data-entry system, it might take five minutes to transcribe the data on a page. But one page typically gives us information on one day in one place. We need to reconstruct the weather everywhere in the world for as far back in time as we have records in the archives – we need to transcribe all of the pages (from all of the archives), and there are many millions of them. If one page takes five minutes, 10 million pages is about 400 person-years of full-time work. That is a lot of people, or a lot of time, and we cannot afford either. How can we go faster?

One possibility is to ask for help: There is a long tradition of 'citizen science' – large groups of amateurs volunteering their time and expertise to research projects, and we have run several citizen science projects, mostly using the Zooniverse platform. These projects have been very successful – in some cases we have had contributions from more than 10,000 volunteers – and as well as transcribing the weather data on pages, volunteer participants have been particularly successful in recording and cataloguing the history and marginal comments in the documents.

But citizen science, while effective, is limited in scale. Projects with 10,000 or 100,000 pages to read and transcribe can work very well. But it is not clear how to scale up to 10 million pages. So, we are currently exploring a second possibility – we are looking to artificial intelligence (AI).

Multimodal large language models – such as the well-known ChatGPT – are very promising tools for document analysis. They can read and interpret images, and they cope very well with variations in image format – so they are not blocked by the many different formats of documents in the archives. We are still investigating how best to use it, but this rapidly developing AI technology offers enormous potential to turn archival records into scientific data.

Once the data have been transcribed, they must then go through a rigorous process of quality control. Individual pieces of data must be consistent with each other, meaning that together they produce a picture that conforms to the basic laws of physics. Original paper records can contain errors, made by the original writer, for instance in the transposition of numbers. Quality control can either correct these original mistakes or discard the data entirely. The document transcription process also captures metadata where this is present in the original document. Metadata might, for instance, record the type of observing instrument and its exposure or position, the altitude of a terrestrial observation or a change of instruments or location; all information that might explain an otherwise anomalous piece of data. Only after these processes are completed are the data from the original paper documents in a suitable form for use in climate reconstructions and reanalyses.

A Norwegian Contribution to Climate Science

A new study recently published in *Geoscience Data Journal* presents a collection of previously unavailable data for the period of 1929–1940 recovered from various accounts originally written in Norwegian. These documents were associated with whaling vessels of Norwegian whaling companies as well as accounts of vessels from the companies with a UK or Irish registration that were operated by crews of Norwegian origin. Since whaling was essentially an international business the documents are now hosted by various national and academic archives across the world. A backbone of this study is a collection of logbooks from the whaling companies A/S Thor Dahl and Pelagos A/S from Tønsberg, Norway. These are archived in the Vestfold archive and the Whaling Museum in Sandefjord, Norway, which in the twentieth century became a major centre of the Norwegian whaling industry. However, several relevant documents for this study were found outside Norway too, such as the accounts of whaling companies Hektoria Ltd., Star Whaling Co. Ltd. and United Whalers Ltd. of London, UK, as well as The South Georgia Co. Ltd. and Sevilla Whaling Co. Ltd. of Dublin, Ireland. Despite their association with British/Irish companies, these accounts were, with a few exceptions, of Norwegian layout/standard and written in Norwegian, as vessel crews would be hired in Norway. This collection of documents is presently stored in the archive of the Sea Mammal Research Unit (SMRU) at the University of St. Andrews in the UK. A few relevant documents were also recovered in the library of the Norwegian Polar Institute, Tromsø, Norway.

With about 50 documents transcribed and analysed already in the first phase of the project, the resulting dataset set comprises some 8000 unique weather records and 4000 notes on sea ice from austral summers for the study period, representing an important contribution to the ongoing effort on historical weather data recovery. Moreover, the data were digitised and formatted in a way that simplifies the data ingestion by major climate data archives and ensures their future use in research on climate and environment.

The European Perspective

Europe has a long history of instrumental measurements, reaching back to the seventeenth century (Figure 2), with the first coordinated data collection activities since the 1710s. Europe also has a long history of rescuing historical observations and of using historical measurements in climate change research. This includes Scandinavia, which has some of the longest records (Bergström & Moberg, 2002; Hestmark & Nordli, 2016; Moberg, 1998). Despite these efforts, a lot remains to be done. Data-rescue activities are often oriented towards specific needs. To better demonstrate climate change, long temperature records are targeted. Dynamical reanalyses use sub-daily pressure, generating the need for pressure data (including short records). Perhaps future AI-related methods will be able to also use non-instrumental observations, such as cloud cover or general weather conditions, leading scientists back to the archives to digitise these. Changes in research questions and in technology revalue archive work.

In addition to instrumental measurements, documentary data can also provide precise climate information. Specifically, ice phenology (dates of freezing and thawing of rivers or harbours) and plant phenology (Norrgård & Helama, 2002) provide quantitative climate information over the past 300 years.

How Can Archives Contribute to the Global Scientific Effort? Types of Documentation and Data Needed

Every state and regional archive, museum, university and library has the potential to hold valuable weather and climate records. These records could be a journal or diary, a

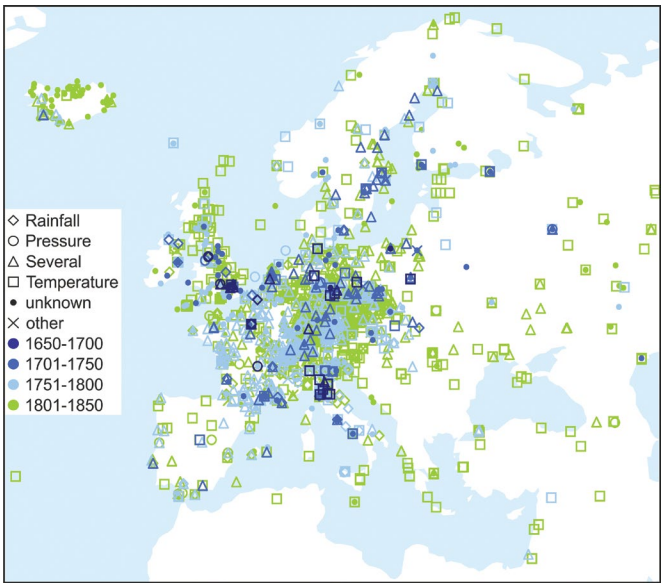

Figure 2. Early instrumental series from Europe as a function of variable and start year from a recent inventory (Brönnimann et al., 2019).

lighthouse register, a formal document such as a ship's logbook, or a meteorological register kept by a weather service or private individual. However other less likely documents also contain valuable weather information, such as insurance and agricultural records, ecclesiastical and missionary records, colonial records and newspapers, to name just a few. Climate scientists can use data from records of snow and ice depth, from rainfall, and descriptions of storms and the damage that can often result from extreme weather. More valuable still are instrumental records of barometric pressure, air and sea temperatures, wind speed and direction.

Working with Archives

Above, it was noted that archives and archivists are critical for the success in gathering more historical weather data from the past. This point cannot be overemphasised. Data-rescue experts are very few in number and even the very modest funding needed for data-rescue activities is always difficult to obtain. The reason for this is that data rescue is not seen as producing immediate or marketable scientific results. It is, however, a critical component for obtaining new climate data, but funders often fail to see the value of such mundane activities. We may deplore this situation, but this is the reality we live with, so other avenues must be explored. As well as the need for better funding, most of the global expertise for data-rescue activities lies with a small and dwindling number of individuals. There are thousands or archives around the world that need to be explored for climate data, and their collections digitised, so even with more funding, the available human resources are limited. Therefore, we need to rethink the data-rescue process, and this is where Norwegian archivists can begin to help the global science community.

The needs of the climate science community make archives particularly relevant to a range of pressing global problems connected with environmental change. This raises and transforms dusty old records into something of immediate relevance, giving the archives an enhanced importance that has an appeal both to the public and to funding bodies. Working and cooperating with scientists is a winning formula for everyone. Nevertheless, there are too few expert individuals to carry out the essential first steps of exploring the archive

collections for weather data. One way that archivists could help the scientific community is to advise of relevant or potentially relevant collections. This might be done through existing knowledge of their collections or by the recruitment of local volunteers of 'citizen scientists' to seek out and document new data sources. This opens a door for many exciting opportunities for public outreach.

Norway has a rich archival heritage that has already proven its worth to the climate science community, but there is so much more that can be done, and needs to be done. We actively encourage archives to contact us to work out future programmes of collaboration and cooperation. If you want to help, or think that you may be able to assist us, then please contact one of the authors.

References

Bergström, H., & Moberg, A. (2002). Daily air temperature and pressure series for Uppsala (1722–1998). *Climatic Change*, 53, 213–252.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538.

Brönnimann, S., Allan, R., Ashcroft, L., Baer, S Barriendos, M., Brázdil, R., Brugnara, Y., Brunet, M., Brunetti, M., Chimani, B., Cornes, R., Domínguez-Castro, F., Filipiak, J., Founda, D., Herrera, R. G., Gergis, J., Grab, S., Hannak, L., Huhtamaa, H., … Wyszyński, P. (2019). Unlocking pre-1850 instrumental meteorological records: A global inventory. *Bull. Amer. Meteorol. Soc. 100*(12)**,** ES389–ES413 (https://doi.org/10.1175/BAMS-D-19–0040.1).

Burn, M. J., & Palmer, S. E. (2015). Atlantic hurricane activity during the last millennium. Scientific Reports, 5. https://doi.org/10.1038/srep12838.

Chand, S. S., Walsh, K. J. E., Camargo, S. J., Kossin, J. P., Tory, K. J., Wehner, M. F., Chan, J. C. L., Klotzbach, P. J., Dowdy, A. J., Bell, S. S., Ramsay, H. A., & Murakami, H. (2022). Declining tropical cyclone frequency under global warming. *Nat. Clim. Chang.* 12, 655–661. https://doi.org/10.1038/s41558–022–01388–4

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., … Worley, S. J. (2011). The Twentieth Century Reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, *137*(654), 1–28.

Donat, M. G., Alexander, L. V., Herold, N., & Dittus, A. J. (2016). Temperature and precipitation extremes in century-long gridded observations, reanalyses, and atmospheric model simulations. *Journal of Geophysical Research: Atmospheres*, *121*(19), 11174–11189. https://doi. org/10.1002/2016JD025480.

Hestmark, G., & Nordli, Ø. (2016). Jens Esmark's Christiania (Oslo) meteorological observations 1816–1838: The first long-term continuous temperature record from the Norwegian capital homogenized and analysed. *Climate of the Past*, *12*(11), 2087–2106

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., & Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1421.

Moberg, A. (1998). Meteorological observations in Sweden made before A. D. 1860. *Paläoklimaforschung*, 23, 99–119.

Norrgård, S., & Helama, S. (2002). Tricentennial trends in spring ice break-ups on three rivers in northern Europe. *The Cryosphere*, *16*(7), 2881–2898.

Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K., & Anandkumar, A. (2022). Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv preprint arXiv:2202.11214.

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., Allan, R., Yin, X., Vose, R., Titchner, H., Kennedy, J., Spencer, L. J., Ashcroft, L., Brönnimann, S., Brunet, M., Camuffo, D.,

Cornes, R., Cram, T. A., Crouthamel, R., Domínguez-Castro, F., … Jones, P. D. (2019). Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Q J R Meteorol Soc*. 2019; 145: 2876–2908. https://doi.org/10.1002/qj.3598

Wilczak, J. M., Akish, E., Capotondi, A., & Compo, G. P. (2024). Evaluation and Bias Correction of the ERA5 Reanalysis over the United States for Wind and Solar Energy Applications. *Energies* 2024, 17, 1667. https://doi.org/10.3390/en17071667