USING EXPIRED WEATHER FORECASTS TO SUPPLY 10 000Y OF DATA FOR ACCURATE PLANNING OF A RENEWABLE EUROPEAN ENERGY SYSTEM

Petr Dolezal
Department of Computer Science
University of Cambridge
pd423@cst.cam.ac.uk

Emily Shuckburgh
University of Cambridge
emily.shuckburgh@cst.cam.ac.uk

ABSTRACT

Expanding renewable energy generation and electrifying heating to address climate change will heighten the exposure of our power systems to the variability of weather. Planning and assessing these future systems typically lean on past weather data. We spotlight the pitfalls of this approach—chiefly its reliance on what we claim is a limited weather record—and propose a novel approach: to evaluate these systems on two orders of magnitude more weather scenarios. By repurposing past ensemble weather predictions, we not only drastically expand the known weather distribution—notably its extreme tails—for traditional power system modeling but also unveil its potential to enable data-intensive self-supervised, diffusion-based and optimization ML techniques. Building on our methodology, we introduce a **dataset** collected from ECMWF ENS forecasts, encompassing power-system relevant variables over Europe, and detail the intricate process behind its assembly.

1 INTRODUCTION AND BACKGROUND

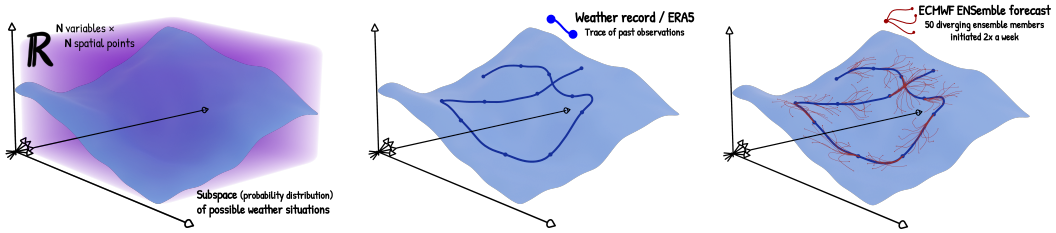
To address climate change we need to electrify large parts of our society and decarbonize the electricity grid. With an uncertain future of nuclear power in Europe [1] and the US [2], it can be expected that a large portion of electricity generation will come from the other source of carbon-free electricity: renewable technologies including photovoltaics, on-shore and off-shore wind turbines and hydro-power. Together with a much larger temperature dependency of electricity demand, as the electrification of heating goes underway, the behaviour of these future power systems will be largely driven by weather [3].

1.1 WEATHER IN POWER SYSTEMS PLANNING

Traditionally, power system planning didn't examine weather since the power generation was controllable and responded to demand. Later, the growing seasonal signal in load and other parts of the system, begun to be supplied through a *representative year*. This practice is still prevalent in power system engineering, as evidenced by the latest guidelines of the European transmission coordinator for their ten year development cycle, which considered 30 years of weather but ultimately only chose 3 specific years (1995, 2008 and 2009) to act as the *climate representative years* [4, p. 15]. Studies focusing on renewable generation usually consider several years of historical data, often taken from reanalyses [5, 6], which assimilate observed data onto easy-to-use grids.

It's crucial to recognize that, as we're designing and planning weather-dependent systems for the next 5-20 years, we simply don't know what the weather will be. We have to treat future weather as a probability distribution of possible scenarios, in which our power system still needs to function.

As shown above, this distribution is often implicitly constructed directly from the existing weather record. When considering a small scale study (such as a single wind turbine), then the weather record can provide a reasonably detailed histogram of the conditions that can occur, however, in continental-scale power-system studies considering many aspects of the weather, the input space



(a) (b) (c)

Figure 1: An illustration of the input space of a weather-driven model. Purple volume represents all possible inputs to the model, blue sheet the subspace/distribution of plausible physical scenarios, and the blue and red lines different ways to estimate this subspace. The weather record (blue) forms a single continuous trace but covers the distribution sparsely. The Ensemble forecasts (red) are regularly initiated from the weather record, but due to chaos in the system, they diverge and predict different plausible weather situations. Due to their numbers the ensemble predictions provide a much more thorough coverage.

grows exponentially (Fig. 1a) and the existing 40 or 80* years of weather record isn't enough to cover the variability of the weather system, including all its spatial correlations (Fig. 1b).

By its definition, climate change also shifts of this distribution, leaving older weather records potentially less informative of the future. While global climate/circulation models (GCMs) aim to project future weather patterns, they had prioritized different objectives, resulting in granularity that often falls short of what's needed here. This gap between energy system and climate modeling has been brought front and centre by Craig et al. [9],

The approach we present offers a powerful compromise. By leveraging the past ("expired") *ensemble weather forecasts*, we build the distribution not just based on what has happened, but also what could have happened (Fig. 1c). We can provide over 10 000 years of independent weather scenarios with higher fidelity than GCMs; and, although derived from past conditions, since they cover the last two decades, these forecasts reflect an atmosphere already influenced by anthropogenic warming [10, p. 42].

1.2 Ensemble forecasts by the European Centre for Medium-Range Weather Forecasts (ECMWF)

The atmospheric weather system is famously a chaotic system, which means even though we can derive deterministic physical laws governing its constituent parts (e.g. differential equations), if we use these laws to predict the future development of the system, any small discrepancy in initial conditions between two scenarios will keep growing until the two predictions appear completely unrelated from each other. This is popularly known as the butterfly effect.

In weather forecasting, this discrepancy will always be present, not just because through our observations we obtain a limited understanding of the state of the atmosphere, but because we introduce further discrepancies during the prediction (from an imperfect knowledge of the physical laws, or the finite resolution of e.g. our numerical integration).

Ensemble weather forecasts are one way to address this challenge: instead of performing one numerical integration and taking that to be our forecast, we perform many (an ensemble), each with slight perturbations in the initial conditions or the equations used for the prediction, within a range that can conceivably match our ignorance of the system. These forecasts can thus provide not just predictions of the weather but also our uncertainty in any prediction, based on the variance between the ensemble members.

ECMWF produces many forecasts, but crucially since 2015, twice a week, it has run an *ENS Extended*, a global 51-member ensemble forecast, integrated forward for 6 weeks (1104h)[11]. Since 2016, it has run a companion product to ENS Extended, called a *Reforecast* or *Hindcast*, where it used

*The dominant dataset, ERA5 [5] has been recently back-extended from 1979 to 1940 [7, 8].

the same latest version of their physical model, but instead of initiating it on the current state of the atmosphere, it used initial conditions from the same date but every year in the past 20 years and integrated 11 ensembles in each. This practise allows ECMWF to compare those reforecast predictions to the weather then recorded and consequently understand the strengths and weaknesses of their current model and provide constraints on the active forecast [12].

1.3 Machine Learning applications

Optimization On reason why power systems planners used *representative years* is due to the computationally intensive nature of the Optimum Power Flow problem. Bypassing this computational cost with machine learning algorithms is a hot topic of research, but these algorithms are often trained on IEEE abstract test grids with synthetic data [13]. Small scale in-the-loop optimization demonstrations such as Huang and Chen [14] show success but require 12500 synthetic training samples and larger systems can be expected to require even more. Unlike ERA5, this dataset can supply such breadth of data from the real world.

Diffusion In some of the latest climate simulation downscaling methods, a diffusion process capable of generating novel samples is trained on high resolution weather scenarios, and then the new samples generated are conditioned to match outputs of the coarse climate model [15]. Larger sample of possible weather situations might allow to train much better and more expressive diffusion processes.

2 Data Collection and Archival process

The dataset presented has been collected from previously run ECMWF numerical integrations. Continually assimilating the latest observations, ECMWF reruns the full integration regularly and provides easy access to this operational forecast to its member states and paying customers. After the predictions are superseded by the next integration, all the previous outputs are, still kept, but moved to an offline tape archive. This data is then governed by the Creative Commons Attribution 4.0 International (CC BY 4.0) license. To access the MARS Archive a license must have been acquired from the Met Office (UK representative).

Reading anything from the tapes is subject to a queue of requests and therefore a slow process. Each integration run is stored on a separate tape (reforecasts on multiple), so a single forecast request takes 4 hours to proceed (22 hours for multi-tape reforecast requests). Due to the maximum request size limitation and concerns about the storage of the dataset, only a set of the most relevant variables has been selected (Table 1), the spatial extent was cropped to a lon-lat rectangle (N/W/S/E = 72/-12/33/34.8) covering the European grid and its surrounding areas. The data has been interpolated from the native spectral representation on an octahedral grid to a uniform 0.4° or 0.5° lat/lon grid. A script created to continuously submit the requests and the requests used are available in appendix A.3.

3 Evaluation and coverage

Ensemble independence The divergence required to ensure the weather samples are indeed independent of their initial conditions, has been evaluated through variance between ensemble members. In figure 2, it can be seen that past two weeks, the variance remains at its maximum value, therefore there is 31 days or 4.5 weeks of essentially independent data.

Size of the dataset Since we can obtain 4.5×51 weeks of data from every forecast and $4.5\times20\times11$ from every reforecast and they've been run twice a week, per year of the model runs we obtain 460 and 1960 years of data. This is fully available for the years 2016-2022 providing more than 14,000 years of data in total.

Extended distribution To illustrate the power of this dataset, figure 3 contains the comparison of a small subset of ECMWF forecast runs to ERA5 scenarios. A single forecast contains a broader range of scenarios that can occur in the wind power system than a full year of ERA5 scenarios.

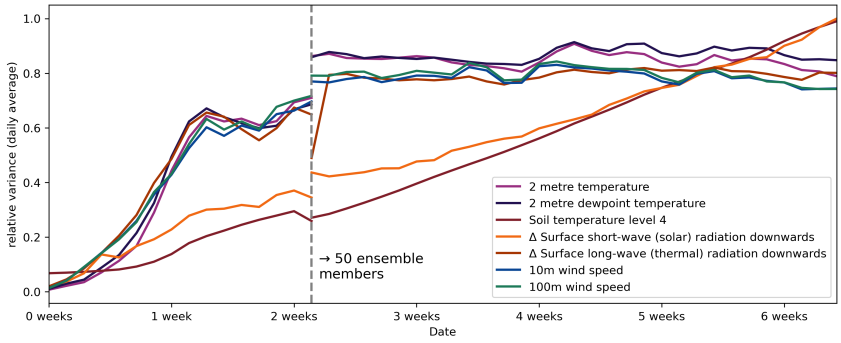


Figure 2: Variance between ensemble members relative to the maximum for each variable. The seasonal signal in winter solar radiation doesn't allow a direct comparison.

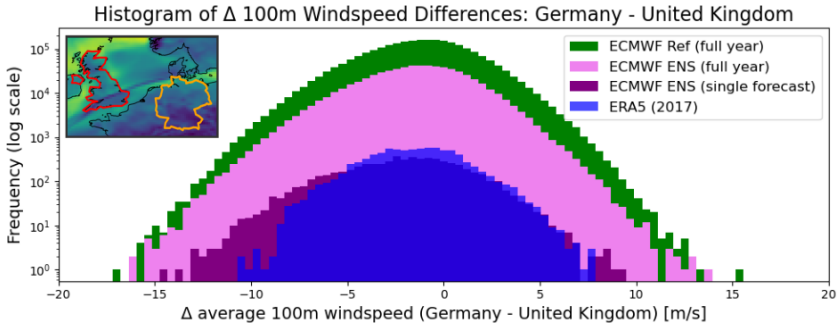


Figure 3: Histogram of difference between 100m wind speed averaged over the spatial extent of two countries (Germany - United Kingdom, shown in the inset). Histograms from four different sources of data are overlaid, all sampled 6 hourly: a full year of ERA5 reanalysis, a single (winter) 4-week 51-ensemble ENS forecast, an accumulation of ENS forecasts over the whole year, and accumulation of Reforecasts over a whole year (which contain data corresponding to the 20 years prior). The y-axis representing counts is scaled logarithmically, showing the much wider distribution present in the ECMWF dataset.

Limitations For both the Extended forecast and the Reforecast, ECMWF only saves the state of the numerical system 6 hourly. This means this dataset can't be used directly to evaluate immediate dynamic responses of the system, such as power ramping.

Future development Since June 2023, the ENS Extended is run daily with 100 members, increasing the multiplicative factor to 3150 years/year once a whole year of forecasts is computed.

4 Conclusion

The ever-growing threat of climate change demands innovative approaches to planning and assessing future power, or any other weather-driven systems. This study underlines the limitations of using past weather data for such endeavors and promotes the use of an extensive, varied weather scenario bank, available through our dataset. This proposed methodology not only provides a comprehensive picture of possible weather events—including the extreme outliers—but also paves the way for novel machine learning applications and their scaling, that couldn't have been imagined in the past.

Acknowledgements

This research was funded by the UK Research and Innovation (UKRI) Centre for Doctoral Training (CDT) AI for the research of Environmental Risks (AI4ER), computing resources were provided by

the University of Cambridge. The data was generated by the European Centre for Medium-Range Weather Forecasts (ECMWF) and procured freely with a MetOffice UK license.

REFERENCES

[1] Marco Sonnberger et al. "Climate concerned but anti-nuclear: Exploring (dis)approval of nuclear energy in four European countries". In: *Energy Research & Social Science* (May 2021). DOI: 10.1016/j.erss.2021.102008.

[2] A. Abdulla et al. "Limits to Deployment of Nuclear Power for Decarbonization: Insights from Public Opinion". In: *Energy Policy* (June 2019). DOI: 10.1016/j.enpol.2019.03.039.

[3] ENTSO-E. *System Needs Study: System dynamic and operational challenges*. Ten-Year Network Development Plan 2022. May 2023. URL: https://tyndp.entsoe.eu/resources/system-dynamic-and-operational-challenges.

[4] ENTSO-E. *Implementation Guidelines for TYNDP 2024*. Version Draft for public consultation. Sept. 2023. URL: https://tyndp.entsoe.eu/resources/tyndp-2024-implementation-guidelines.

[5] Hans Hersbach et al. "The ERA5 global reanalysis". In: *Quarterly Journal of the Royal Meteorological Society* 146.730 (2020), pp. 1999–2049.

[6] Ronald Gelaro et al. "The modern-era retrospective analysis for research and applications, version 2 (MERRA-2)". In: *Journal of climate* 30.14 (2017), pp. 5419–5454.

[7] Bill Bell et al. "The ERA5 Global Reanalysis: Preliminary Extension to 1950". In: *Quarterly Journal of the Royal Meteorological Society* 147.741 (2021), pp. 4186–4227. ISSN: 1477-870X. DOI: 10.1002/qj.4174. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.4174 (visited on 09/29/2023).

[8] Jose Andres Perez Leon. *ERA5 reanalysis now available from 1940*. en. Text. Apr. 2023. URL: https://www.ecmwf.int/en/newsletter/175/news/era5-reanalysis-now-available-1940.

[9] Michael T. Craig et al. "Overcoming the disconnect between energy system and climate modeling". In: *Joule* 6.7 (2022), pp. 1405–1417. ISSN: 2542-4351. DOI: https://doi.org/10.1016/j.joule.2022.05.010.

[10] *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II, and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Tech. rep. Specific page citation, e.g., p. 50, can be mentioned here if needed. Geneva, Switzerland: Intergovernmental Panel on Climate Change, July 2023, pp. 35–115. DOI: 10.59327/IPCC/AR6-9789291691647. URL: https://doi.org/10.59327/ipcc/ar6-9789291691647.

[11] Frédéric Vitart et al. *Extended-range prediction*. eng. Nov. 2019. DOI: 10.21957/pdivp3t9m. URL: https://www.ecmwf.int/node/19286.

[12] T Gneiting. *Calibration of medium-range weather forecasts*. eng. Mar. 2014. DOI: 10.21957/8xna7glta. URL: https://www.ecmwf.int/node/9607.

[13] Bin Huang and Jianhui Wang. "Applications of physics-informed neural networks in power systems-a review". In: *IEEE Transactions on Power Systems* 38.1 (2022), pp. 572–588.

[14] Wanjun Huang and Minghua Chen. "DeepOPF-NGT: A Fast Unsupervised Learning Approach for Solving AC-OPF Problems without Ground Truth". en-US. In: *Climate Change AI*. Climate Change AI, July 2021. URL: https://www.climatechange.ai/papers/icml2021/18.

[15] Zhong Yi Wan et al. *Debias Coarsely, Sample Conditionally: Statistical Downscaling through Optimal Transport and Probabilistic Diffusion Models*. 2023. arXiv: 2305.15618 [cs.LG].

A Details of the Dataset

A.1 Parameters of the dataset

The dataset is stored on files indexed by the date of the forecast run that generated it (Mon or Thu). There are several types of files containing different ensemble and timestamp combinations:

Type **d**	50 ensembles of the ENS Extended (post-divergence)	$360 \overset{6h}{\ldots} 1104h$
Type **e**	5 ensembles of the ENS (pre-divergence, for evaluation)	$0 \overset{1h}{\ldots} 90 \overset{3h}{\ldots} 144 \overset{6h}{\ldots} 360h$
Type **r**	20y $\times$ 10 ensembles Reforecast runs	$0 \overset{6h}{\ldots} 1104h$
Type **f**	new daily 100 ensemble ENS Extended runs	$0 \overset{6h}{\ldots} 1104h$

Table 1: Units collected in the dataset

var	unit	name
t2m	$^\circ\mathrm{C}$	2 metre temperature
w10	$\mathrm{m\,s^{-1}}$	10m wind speed
w100	$\mathrm{m\,s^{-1}}$	100m wind speed
ssrd	$\mathrm{J\,m^{-2}/h}$	Δ Surface short-wave (solar) radiation downwards
strd	$\mathrm{J\,m^{-2}/h}$	Δ Surface long-wave (thermal) radiation downwards
ssr	$\mathrm{J\,m^{-2}/h}$	Δ Surface net short-wave (solar) radiation
ro	m	Runoff
stl4	$^\circ\mathrm{C}$	Soil temperature level 4
d2m	$^\circ\mathrm{C}$	2 metre dewpoint temperature
sp	Pa	Surface pressure

A.2 Dataset availability

The European dataset as collected is being made available through CEDA Archive (pending approval) at `https://catalogue.ceda.ac.uk/uuid/7783f79c7080456088d98a34ca238bfa`

Table 2: Currently available data containing 3700y of weather

	available period	effective amount of data
Type **d**	2017-18, 2022	1300y
Type **e**	2017-18	N/A
Type **r**	2017	1700y
	06/2023-04/2024	1400y
Type **f**	06/2023-03/2024	2500y

A.3 Dataset Compilation and Utilization Codebases

The tools used for collection of the dataset, including scripts for communication with the ECMWF Archive and specific requests, are openly available at `github.com/elkir/ens_download_manager`.

Tools for working with the datasets, including loading, processing, visualization and distributed processing on SLURM systems, are openly being developed and made available at `github.com/elkir/delorean-datasets`.

A.4 Examples of data present in the dataset

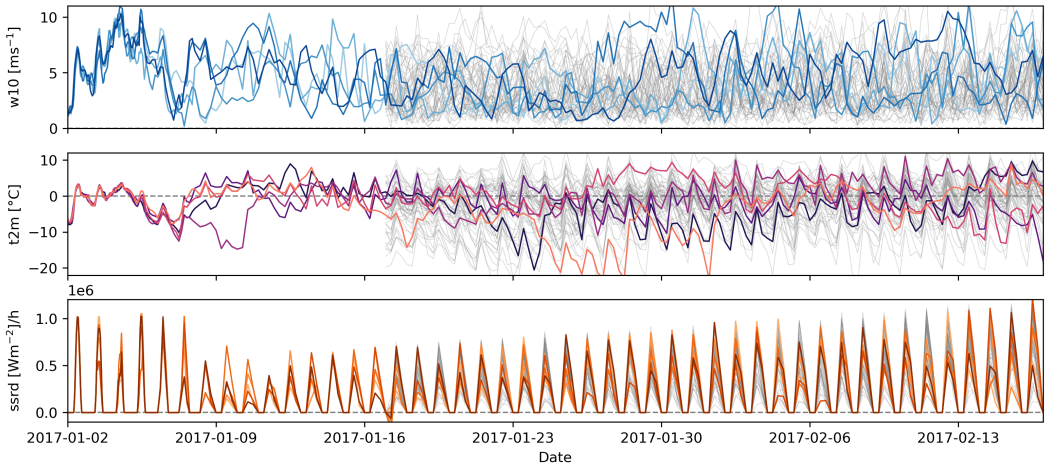


Figure 4: An example of a single ECMWF ENS Extended run prediction for Vienna. The three variables most dominant in the power system are shown. 51 ensembles are run in parallel for 6 weeks and they quickly diverge from initial conditions (all 51 shown only from week 2). The three variables most dominant in the power system are shown - 2m temperature (t2m), surface short-wave (solar) radiation downwards (ssrd) and 10m wind (w10).

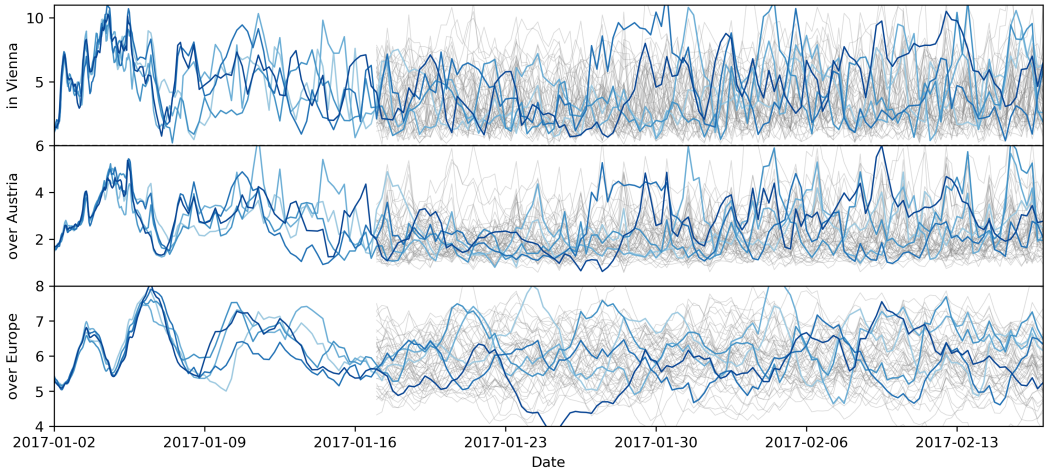


Figure 5: Same ENS forecast as above in Fig. 4. The top graph shows the same 10m wind trace from a single location (Vienna), the bottom two an average over a country and the full spatial extent of the data. It is clear that the ensembles eventually predict differing conditions not just locally but over the whole area, in demonstration of the butterfly effect