scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Baroclinic Wave Simulation Ensemble: a Machine Learning ready dataset

Clément Bouvier[1]✉, Joona Cornér[1], Antti Toropainen[1], Andy Bowery[2], Glenn Carver[3], Sarah Sparrow[2], David Wallom[2] & Victoria Anne Sinclair[1]

A large ensemble of 6,500 different baroclinic wave simulations have been run, processed and provided to study extra tropical cyclones and mid-latitudes dynamics. The data were generated using OpenIFS@home, an open science climate*prediction*.net project allowing the distribution of the computation of the ensemble with the OpenIFS 43R3v2 model. For each simulation, the cyclones were tracked and 89 features -including 16 intensity measures- were extracted. The presented dataset is composed of the raw output of the OpenIFS model for 6,388 of the 6,500 members of the ensemble and the extracted features of the tracked cyclones. The distribution of the minimum mean sea level pressure and the maximum relative vorticity at 850 hPa is plotted to enable comparison with studies that have analysed ETCs in reanalyses and climate model data. The computational failure of the missing 112 ensemble members is statistically assessed and explained. Outside of OpenIFS, the dataset and the associated code and configuration files are available and accessible.

Background & Summary

This dataset is a collection of features extracted from an ensemble of idealised moist baroclinic wave simulations (BWS)[1,2]. Baroclinic waves are synoptic-scale, alternating low and high pressure systems that grow in the midlatitudes due to baroclinic instability. These waves are important parts of the Earth's global circulation as they transport energy polewards. Additionally, BWS can be used to investigate the dynamics of extra tropical cyclones (ETCs).

The BWS initial background states are meteorologically and numerically stable when run without the addition of an unbalanced perturbation, and are controlled by seven entry parameters[3]. Each BWS is initialised with a different set of entry parameters sampled from a Latin Hypercube, representing a different atmospheric initial state and allowing different realistic baroclinic wave development. Then, each developing ETC is tracked and characterised with a set of 89 features. These different feature sets constitute the dataset which is presented here and is Machine Learning ready, allowing exhaustive comparisons of a vast array of different background states and the resulting baroclinic developments using traditional or deep learning methodologies. Moreover, this dataset can be used with standard meteorological methods and as an alternative to re-analysis datasets to study ETCs. This comparison can be extended to understand the intricate relationship between the background atmospheric state and the eventual intensity of a cyclone that develops in it.

The dataset consists of one xml file and two repositories. The first repository contains 6,388 folders, each representing one BWS as simulated by the Open Integrated Forecasting System (OpenIFS) 43R3v2[4]. The raw outputs are compressed and stored in these folders. In total 28 variables are output on either surface levels or on 28 pressure levels at a horizontal resolution of 125 km at the equator and 3 hour temporal resolution. The second repository has 22,259 comma-separated value files (.csv) containing 89 features extracted for each BWS and associated ETC. The features can be classified into four categories: background-related features, track-related features, dynamical intensity measures, and impact-relevant intensity measures.

[1]INAR/Physics, University of Helsinki, Helsinki, 00560, Finland. [2]e-Research Centre, University of Oxford, Oxford, OX1 3PJ, United Kingdom. [3]climateprediction.net, University of Oxford, Oxford, OX1 3PJ, United Kingdom. ✉e-mail: clement.bouvier@helsinki.fi

Input parameter	Units	Min	Max	Data type
n, Jet width	Dimensionless	1	6	integer
b, Jet height	Dimensionless	0.5	2.5	float
u_0, Average zonal wind speed	$m\,s^{-1}$	15.0	75.0	float
T_{v0}, Average virtual temperature	K	265.0	300.0	float
RH_0, Surface level relative humidity	%	0.0	80.0	float
Lapse rate	$K\,km^{-1}$	0.003	0.006	float
Charnock parameter	Dimensionless	0.01	0.035	float

Table 1. Details of the 7 input parameters detailed in the Methods section. Each line correspond to one of the seven input parameter.

Methods

This section describes the methods and infrastructures used to produce the dataset, including the toolboxes used and Application Programming Interface (API).

Moist baroclinic wave simulation. General Circulation Models (GCMs) are the cornerstone of weather forecasting and climate simulation[5]. They can be used to produce operational weather forecasts, predict future climate[6], but also to simulate idealised weather systems enabling specific phenomena, such as convection[7] or baroclinic waves[8] to be studied. Baroclinic waves are synoptic weather phenomena of high and low pressure systems that develop in the mid-latitudes. These waves are fundamental to understand Earth's global circulation as they transport energy and moisture polewards[2,9,10]. To study these phenomena, moist baroclinic wave simulations are performed using the Open Integrated Forecasting System (OpenIFS) cycle 43R3v2[4].

OpenIFS is a version of the Integrated Forecasting System of the European Centre for Medium-Range Weather Forecasts (ECMWF). Cycle 43R3 of IFS was operational at ECMWF from July 2017 to June 2018[4]. The initial background state for these simulations is expressed analytically and setup through the configuration files in an aquaplanet setting. As a result, the jet structure and strength, the average virtual temperature, the surface relative humidity, the lapse rate and the surface roughness can be easily modified. In total, seven input parameters can be controlled to produce different initial background states, and subsequently different baroclinic wave developments. Previous work[3] can be consulted for more details on the background state, the baroclinic wave development and the implementation in OpenIFS 43R3v2. Each baroclinic wave simulation is run with a spatial resolution of T_L159, which corresponds to a specific spectral truncation used to compute the linear terms in the equations of the model[11]. The final grid cell size is approximately 125 km at the equator with 91 sigma levels, and the simulations are global. Alongside the spatial resolution, a model time step of 900 s (15 min), an output frequency of 3 h and a length of simulation of 20 days are chosen. The following physical parameterization schemes have been activated: vertical diffusion (LEVDIF), surface processes (LESURF), large-scale condensation (LECOND), mass-flux convection (LECUMF), prognostic cloud scheme (LEPCLD) and evaporation of precipitation (LEEVAP). The text in parenthesis refers to the OpenIFS namelist option that is set to true to activate each physical parameterization scheme. In addition, the negative humidity fixer (LEQNGT) is also turned on[12]. The radiation schemes have not been triggered. Depending on the initial background state, multiple low and high pressure systems develop. Some initial conditions lead to no development at all. All code and configuration files related to the moist baroclinic wave simulations can be found on this Zenodo repository[13].

Ensemble generation. To generate the ensemble, the seven input parameters are used to define a 7-dimensional hypercube. The input parameters, along with their maximum and minimum values are given in Table 1. A Latin Hypercube Sampling (LHS)[14] is performed in this parameter space to define a list of 6,500 different configurations. The code used to generate the LHS can be found in a GitLab repository[15]. Each configuration results in a specific set of initial conditions and thus a unique baroclinic wave simulation, which was simulated using OpenIFS@home[16]. OpenIFS@home is an open science climate*prediction*.net (CPDN)[17] project allowing the fast computation of the computationally heavy ensemble of OpenIFS forecasts using the Berkeley Open Infrastructure for Network Computing (BOINC)[18] framework. Five days after the start of the ensemble computation, 80% of the members were returned and transferred to the CSC - IT centre for science Ltd[19] infrastructure for further processing as detailed in the next section.

Ensemble processing. All the BWS are processed in order to 1) track the ETCs which develop in the simulations and 2) extract features which further characterise the background state and the developing cyclones. The workflow is presented in Fig. 1 and is fully implemented using the HyperQueue API[20] on the CSC infrastructure[19]. From a list of identification tags (the experiment IDs used by OpenIFS), each BWS is processed independently, between 1 and 4 files are generated and pooled together. The next sections describe in detail the objective cyclone tracking and the feature extraction process.

Objective cyclone tracking. Cyclone tracks are identified with the objective feature tracking software TRACK[21–23]. TRACK uses a Lagrangian approach of tracking individual cyclones by identifying extrema in a given field and following them through time. In order to track developing cyclones, the relative vorticity at 850 hPa (VO-850) at the T_L159 resolution is first truncated to the T42 spectral resolution (310 km at the equator) and the planetary

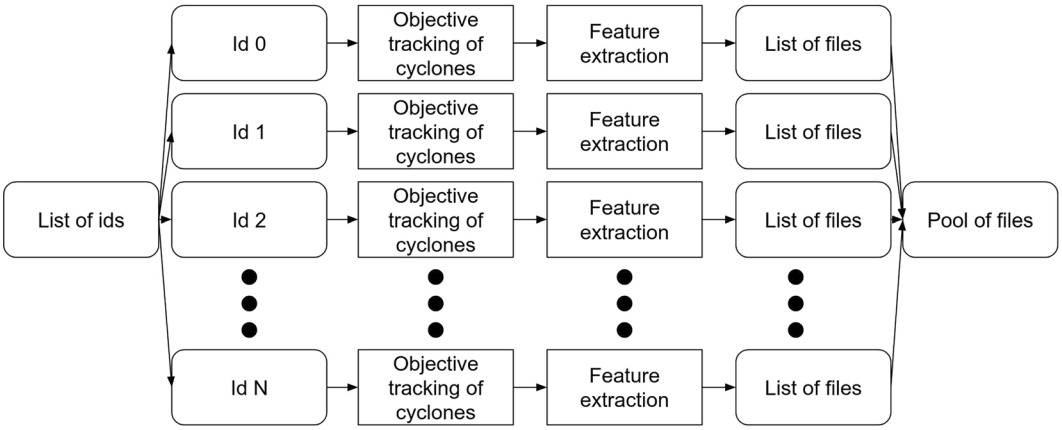

Fig. 1 Distributed workflow for objective cyclone tracking and feature extraction.

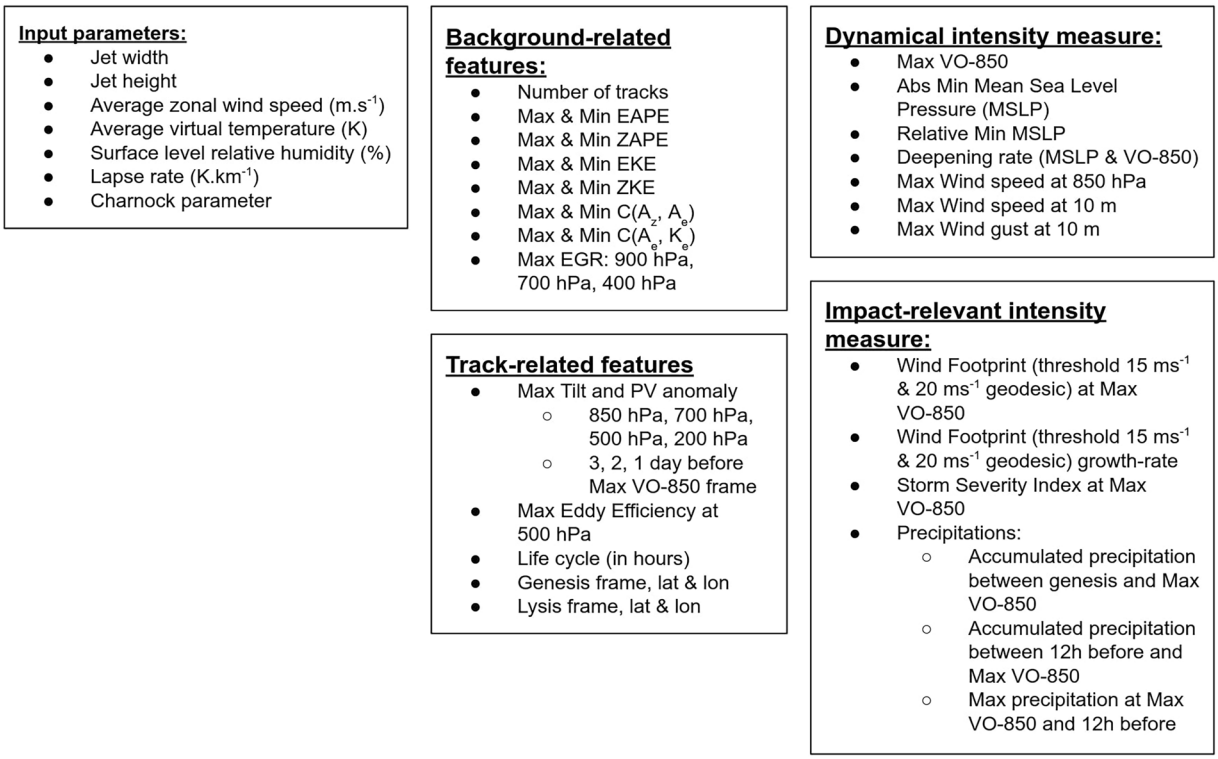

Fig. 2 Taxonomy of the extracted features. The time frame of occurrence of each represented feature is also extracted if not computed at Max VO-850 time.

scale waves (wavenumbers 1-5) are excluded. This truncation ensures that very large- and small-scale features are excluded and only synoptic-scale cyclones are identified. TRACK produces output which consists of the horizontal location (longitude and latitude) and magnitude of the T42 VO-850 maxima for each time frame in each cyclone track. Then, the maximum relative vorticity within 2° geodesic radius is localised using relative vorticity at 700 hPa, 600 hPa, 500 hPa, 400 hPa, 300 hPa and 200 hPa in order to compute the tilt of the cyclone at these different pressure levels with VO-850 as the reference. The tilt is computed iteratively starting from the tracked relative vorticity at 850 hPa. Using the T42 maxima at the next pressure levels (700 hPa), the steepest ascent maximisation within a 5° geodesic radius is estimated using B-spline and the tilt is computed[24]. The tilt is computed alongside the objective tracking[25]. Finally, the cyclone tracks based in VO-850 are filtered to exclude stationary, weak and short-lived systems. Therefore, the tracks need to have a T42 VO of at least $1 \times 10^{-5} s^{-1}$, be at least 1000 km long, and last for at least two days. All TRACK's configuration files can be found in a GitLab repository[15]. The first three cyclones objectively identified in each BWS simulation by TRACK are kept for the feature extraction process.

Number of BWS:	6,388
Data size per BWS:	3.9 GB
Number of compressed files per BWS:	20
Number of uncompressed files per BWS:	485
Spatial resolution:	T_L159 L91
Output frequency and length of simulation:	3 h and 20 days
List of outputted OpenIFS pressure levels (19 levels):	1000 hPa, 950 hPa, 900 hPa, 850 hPa, 800 hPa,
	750 hPa, 700 hPa, 650 hPa, 600 hPa, 550 hPa,
	500 hPa, 450 hPa, 400 hPa, 350 hPa, 300 hPa,
	250 hPa, 200 hPa, 150 hPa, 100 hPa
Number of OpenIFS parameters:	28
List of 3D OpenIFS fields:	Potential vorticity
	Geopotential
	Temperature
	Zonal wind (U)
	Meridional wind (V)
	Specific humidity
	Vertical velocity (W)
	Relative vorticity
	Divergence
	Relative humidity
	Specific cloud liquid water content
	Specific cloud ice water content
	Fraction of cloud cover
List of 2D OpenIFS fields:	Surface pressure
	Logarithm surface pressure
	Sea surface temperature
	Maximum 10 metre wind gust since previous post-processing
	Convective available potential energy
	Maximum 10 metre wind gust in the last 6 hours
	Total column vertically-integrated water vapour
	Large-scale precipitation
	Convective precipitation
	Mean sea level pressure
	10 metre U wind component
	10 metre V wind component
	2 metre temperature
	Evaporation
	Forecast surface roughness
Total size of this repository (compressed raw data):	10.34 TB

Table 2. Data breakdown for each baroclinic wave simulation.

Feature extraction. The feature extraction process can be decomposed into two parts and the features into four categories as represented in Fig. 2. The processing pipeline has been developed using Python.

First, background-related features are extracted. These features are only dependent on the background state and its evolution, meaning that these features are extracted regardless of the development of baroclinic waves. For each BWS, the time-series of the four energies of the Lorenz energy cycle are computed[26,27]: the Zonal mean Available Potential Energy (ZAPE), the Eddy Available Potential Energy (EAPE), the Eddy Kinetic Energy (EKE), and the Zonal mean Kinetic Energy (ZKE). The time-series of conversion terms between the ZAPE and EAPE, and between EAPE and EKE are averaged and extracted between 30° N and 60° N. The details of this calculations are shown in Toropainen (2024)[28]. The absolute maximum, minimum, and their respective time frame of occurrence are extracted. Finally, the maximum zonal mean Eady Growth Rate (EGR) values between 30° N and 60° N at each pressure levels considered (900, 700, 400 hPa) are computed from the potential temperature at a pressure level above and below[29].

Then, the track-related features and the intensity measures are calculated. These features are extracted if at least 1 track is detected during the objective cyclone tracking for a given BWS. From the tracking process, the time frames of Genesis and Lysis are the first and last frame of the track respectively. The life cycle duration is

Number of Features:	89
Number of BWS with 0 baroclinic development:	372
Number of BWS with 1 cyclone:	305
Number of BWS with 2 cyclones:	95
Number of BWS with 3 or more cyclones:	5,248
Total number of files:	22,259
Total size of the repository:	39.3 MB

Table 3. Data breakdown for the feature extraction.

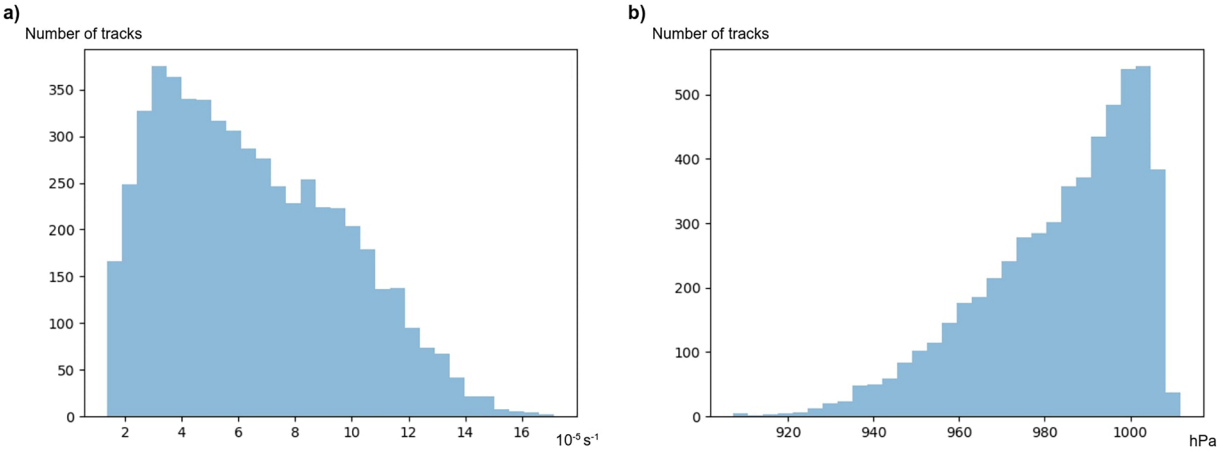

Fig. 3 Total distribution of (**a**) the maximum VO-850 and (**b**) the minimum MSLP. Note that the y-axes differ between the two panels.

the number of frames corresponding to the difference between Genesis and Lysis. The PV anomaly is computed as the difference between the PV field and the PV zonal average at every time frame. Then, the average value within a 2° geodesic radius from the potential vorticity maximum is computed[30,31]. The tilt and PV anomalies are extracted 3 days, 2 days, and 1 day before the frame of maximum tracked VO-850 at all pressure levels considered. The maximum Eddy Efficiency[32] is computed as the maximum of the Eddy Efficiency averaged over a 10° geodesic radius centred on the cyclone's center at 500 hPa.

All intensity measures are computed according to the study by Cornér *et al.*[33] and using the code available in a GitLab repository[15]. A total of 16 intensity measures are computed including maximum vorticity, wind speed (850 hPa, 10 m), wind gust (10 m), the minimum mean sea level pressure (MSLP) and its growth-rate, the wind footprint with 15.0 and 20.0 m s^{-1} threshold and their respective growth-rates, and a Storm Severity Index (SSI)[34]. Finally, three precipitation measures are computed, two accumulated total precipitation (between genesis and the maximum vorticity frame, and the 12 hours before the maximum vorticity frame), and one instantaneous precipitation measure 12 hours before the maximum vorticity frame[35].

Data Records

The dataset is available in FairData.fi[36]. The repository containing the data consists of three parts which are accessible independently. First, the xml used to submit the batch of jobs to the OpenIFS@home infrastructure. Each case is described by a *unique_member_id* and a set of *parameters*. The *unique_member_id* is a set of four walpha-numerical characters unique to a set of *parameters*, it is assigned during the generation of the ensemble and ranges from *a000* to *a50j*. The set of *parameters* corresponds to the seven input parameters used to construct the background state of the baroclinic wave simulations and are presented in Table 1[3]. The parameters are randomly selected and thus there is no meaningful relationship between the *unique_member_id* and the *parameters*.

Secondly, a folder named *batch_1018* which contains the raw output of the OpenIFS@home infrastructure. Each sub-folder corresponds to one of the successful runs (6,388 sub-folder in total, see the Technical Validation section). The *unique_member_id* is included in the name of the sub-folder, which contain a collection of 20 zip archives in which the raw gridded OpenIFS output (GRIB files) is stored. The total size of this dataset is 10.34 TB. A comprehensive description of the *batch_1018* folder is given Table 2.

Lastly, a folder named *ExtractedFeatures* contains the collection of features computed (see Fig. 2) as described in the previous section. There are 22,259 files, each respecting the following nomenclature: *unique_member_id_general* for the background-related features, *unique_member_id_0/1/2* for the track-related features and the intensity measures respectively for the first, second and third baroclinic wave developing in the *unique_member_id* case. A comprehensive description of the *ExtractedFeatures* folder is given in Table 3.

Member ID	n	b	T_{v0}	u_0	RH_0	Laps rate	Charnock parameter
a0jy	6	0.91	297.08	21.04	44.39	0.0052	0.030
a0bs	4	1.57	269.60	42.17	3.14	0.0044	0.027
a1vu	4	1.69	279.77	42.82	55.59	0.0057	0.026
a21u	3	2.025	292.36	68.00	70.79	0.0051	0.032
a333	1	2.31	279.68	56.86	79.94	0.0059	0.019

Table 4. Experiment IDs and input parameters for the 5 cases presented Fig. 5. Number are truncated at 2 significant digits.

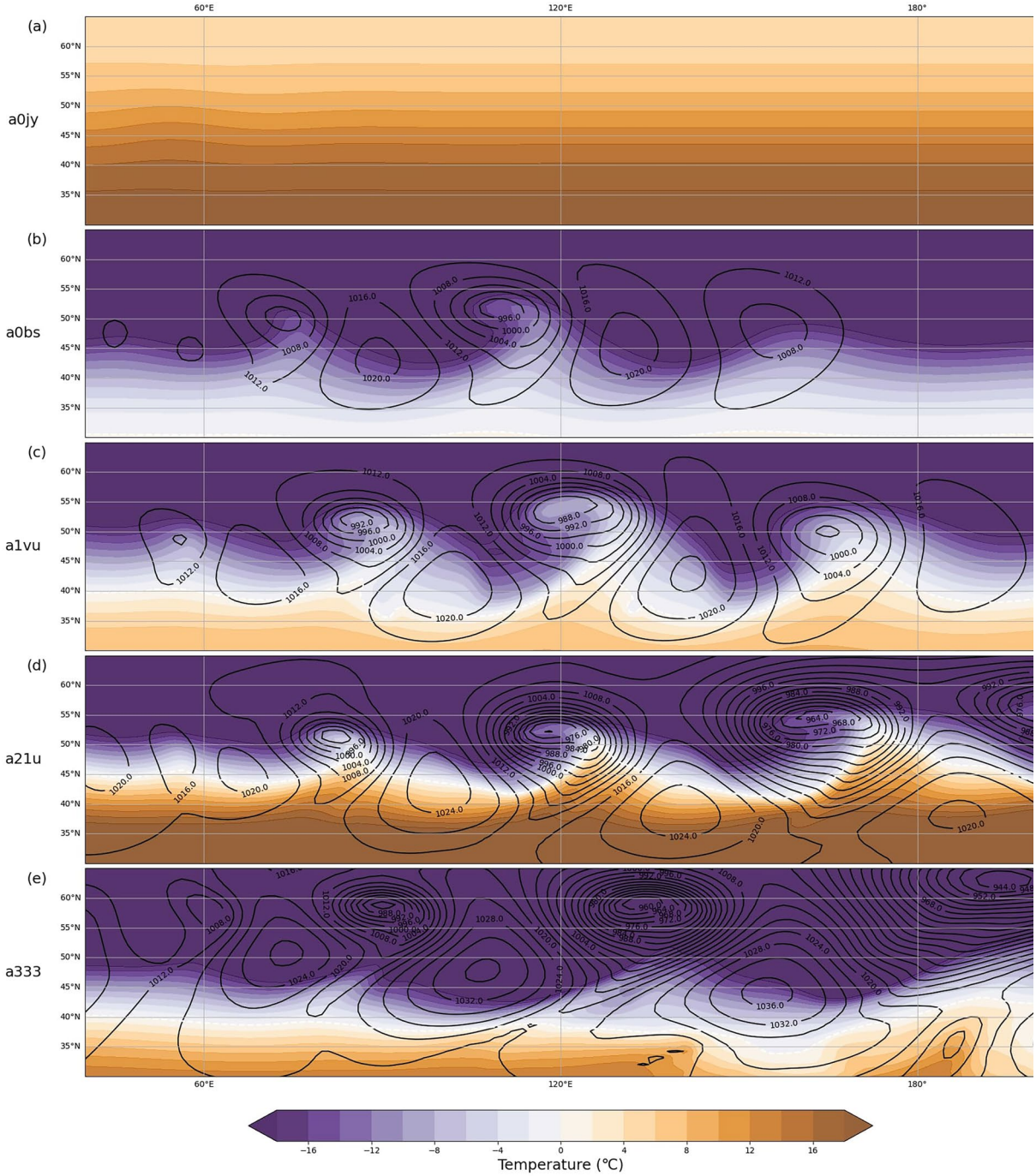

Fig. 4 Baroclinic development for 5 cases at t = 204 hour, from weak to strong waves. The black contours show mean sea level pressure (hPa), and the shading shows the temperature (°C) at 850 hPa. The label on the left correspond to the member ID of the case, their input parameters are given Table 5. Note: this figure does not show the whole (global) model domain; the x axis ranges from 40–220° E, while the y axis ranges from 30–65° N.

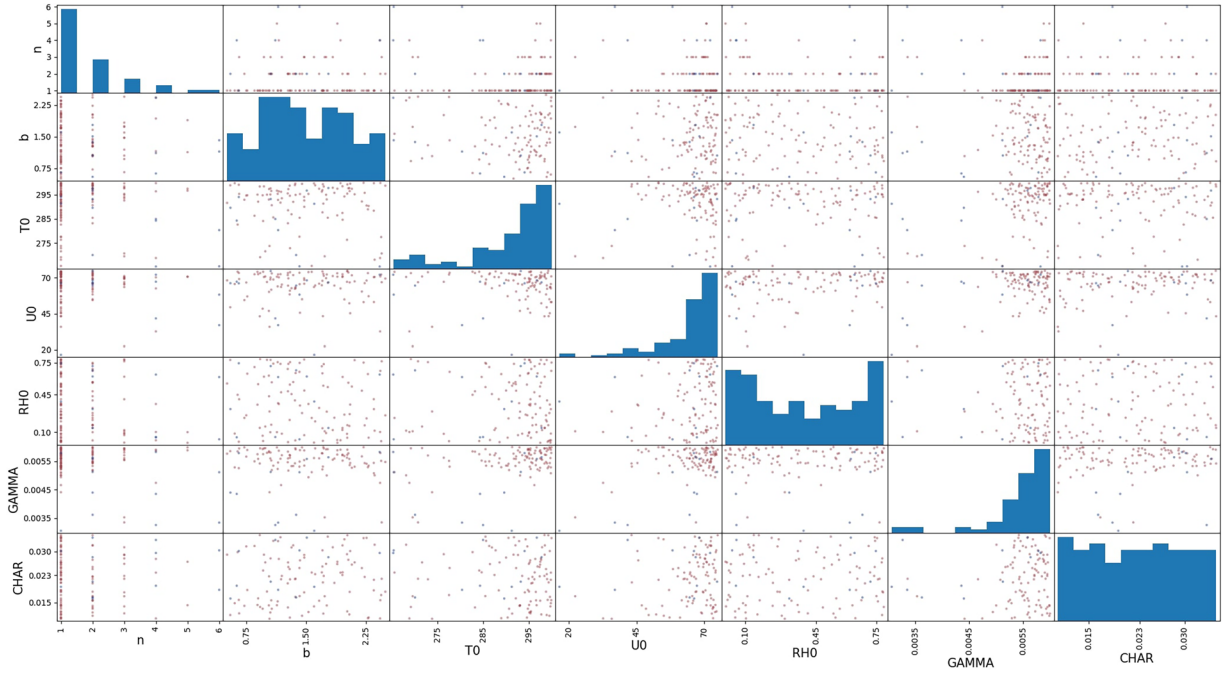

Fig. 5 Projection of missing and hard fail runs in the initial hypercube, brown dots represent hard fail runs, cyan dots the missing runs. The diagonal represents the missing and hard fail distributions of the corresponding input parameter.

Statistical test	n	b	T_{w0}	u_0	RH_0	Laps rate	Charnock parameter
Missing run: U-test p-value	—	—	—	0.014	—	—	—
Missing run: Cramer p-value	—	—	—	$9.98e^{-3}$	—	—	—
Hard fail run: U-test p-value	$1.22e^{-28}$	—	$4.92e^{-19}$	$8.09e^{-32}$	—	$7.77e^{-33}$	—
Hard fail run: Cramer p-value	$1.9e^{-7}$	—	$9.50e^{-11}$	$9.69e^{-10}$	—	$2.73e^{-9}$	—

Table 5. Statistical tests for missing and hard fail runs, p-value are written if both tests are below the level of confidence ($\alpha = 5\%$).

Technical Validation

Of the 6,500 members of the ensemble, 6,388 members have been processed successfully. Of the successful runs 80% have been processed within 5 days of the ensemble being launched on the OpenIFS@home infrastructure[16]. The remaining 20% of the successful runs have been returned within one month of launching the ensemble. OpenIFS@home is an open science and distributed infrastructure meaning that it is dependent on a higher number of technical and human parameters than a traditional High-Performance Computing (HPC) setup, which explains the late return of some of the successful cases. The remaining 112 cases have not been returned (12) or have failed (100). These failures are caused by unrealistically strong cyclones developing which resulted in exceptionally strong updraft winds, causing the OpenIFS 43R3v2 model to become numerically unstable and crash. The 112 unsuccessful runs will be called "hard fail runs" in the rest of the manuscript.

To test the validity of the simulated ETCs, the distribution of the maximum VO-850 and minimum MSLP for the first tracked cyclone in all 6388 successfully processed simulations is plotted in Fig. 3. The two distributions are skewed towards the most intense values, which is similar to the distributions for ETCs found in previous studies which analysed ETCs in the historical climate using reanalysis datasets and in the future climate using climate model simulations[25,33,37–39]. The baroclinic wave simulations[3] therefore result in ETCs with reasonable intensity measure distributions compared to current or projected future climates. Furthermore, a visual inspection of relevant meteorological variables plotted on a map reveal features, such as cold fronts and warm sectors, which resemble those found in analyses from satellite images, reanalyses or model output. Hence we conclude that the ETCs in these idealised simulations resemble ETCs found in the real world. However, the set of input parameters strongly influence the speed and strength of the ETCs development as shown in Table 4 and Fig. 4 for five arbitrarily chosen cases. After 8.5 days of the development (the time step shown in Fig. 4) different combinations of input parameters have resulted in a varying number of mature low pressure systems of varying intensities. Cases with more and deeper low pressure systems undergo more rapid and intense development than those with fewer and shallower ones. For example, from a narrow, low and weak jet in Fig. 4a there is no baroclinic wave development at all. The meridional temperature gradient in this case is weak indicating little baroclinicity which results in slow development of baroclinic waves, as theoretically expected from the Eady

model[40]. In another case with a wide, high and strong jet, and thus large baroclinicity, (Fig. 4e) many deep ETCs associated with large pressure gradients and frontal features can be seen. As a result, the dataset is considered of interest for the study of current and alternative climates. Future work will include comparison with CMIP6[6] projections and in-depth comparison with ERA5[33,41] tracked cyclones.

Figure 5 presents a projection of the missing and hard fail run (112 runs total) in the 7-dimensions hypercube. On the diagonal, the total distribution of the runs is represented. These distributions are compared to the original uniform distributions in order to assess the dependency of the missing or hard fails runs on the seven entry features (see Table 1) with the Mann-Whitney U-test and the Cramér-von Mises test. To consider that an entry feature is increasing the probability of a missing or hard fail run, both tests have to have a p-value below the confidence level. The confidence level is set at 5% for both tests. The results of the statistical tests are presented in Table 5. The missing run distributions depend on u_0, but due to the low sample size for the missing runs, no conclusion can be reasonably drawn for these 12 cases. Concurrently, n, T_{v0}, u_0 and the lapse rate values (see Table 1) increase the probability to have a hard fail run. The hard fail runs are due to the unrealistic background states which can be generated by our implementation[3]. High lapse rate (greater than 0.005 K km^{-1}), initial virtual temperature (superior to 295 K), wind speed (superior to 60 m s^{-1}), with a wide jet stream (small n) create extreme initial conditions, making the OpenIFS 43R3v2 model to become numerically unstable and to crash.

Usage Notes

To manipulate the outputted GRIB files by OpenIFS@home in the folder *batch_1018*, python scripts can be found in a Zenodo repository in the *plotting_scripts* folder[13]. The script *Usage_Script.py* uses the xml file and the *ExtractedFeatures* folder to produce the Figs. 3 and 5, and the statistical tests presented in Table 5. By modifying the beginning of the script, each extracted feature can be filtered and / or plotted. The *Usage_Script.py* is available in a GitLab repository[15].

Data availability

The modified subroutines of OpenIFS, a standalone version to compute the initial zonally uniform fields for wind, the temperature and the geopotential as detailed in Bouvier *et al.* [3] are available in a Zenodo repository[13]. The entire dataset is available in FairData.fi[36]. The dataset can be access *via* the Metax API[42].

Code availability

The licence for using the OpenIFS CY43R3 model can be requested from ECMWF user support (openifs-support@ecmwf.int) and is given free of charge to any academic or research institute. The modified subroutines of OpenIFS, a standalone version to compute the initial zonally uniform fields for wind, the temperature and the geopotential as detailed in Bouvier *et al.* [3] are available in a Zenodo repository[13]. OpenIFS43r3 with the specific changes required to reproduce this work is available on the CPDN platform as OpenIFS@Home, access to which can be obtained via[43]. A GitLab repository consists of the namelists for TRACK[21–23], the feature extractor code to produce the *ExtractedFeatures* folder, the Latin Hypercube Sampling code, and the *Usage_Script.py*[15]. The entire dataset is available in FairData.fi[36]. The dataset can be access *via* the Metax API[42].

Received: 16 May 2025; Accepted: 2 October 2025;
Published online: 18 November 2025

References

1. Hoskins, B. J. & Simmons, A. J. A multi-layer spectral model and the semi-implicit method. *Quarterly Journal of the Royal Meteorological Society* **101**, 637–655, https://doi.org/10.1002/qj.49710142918 (1975).
2. Thorncroft, C., Hoskins, B. & McIntyre, M. Two paradigms of baroclinic-wave life-cycle behaviour. *Quarterly Journal of the Royal Meteorological Society* **119**, 17–55, https://doi.org/10.1002/qj.49711950903 (1993).
3. Bouvier, C., van den Broek, D., Ekblom, M. & Sinclair, V. A. Analytical and adaptable initial conditions for dry and moist baroclinic waves in the global hydrostatic model OpenIFS (CY43R3). *Geoscientific Model Development* **17**, 2961–2986, https://doi.org/10.5194/gmd-17-2961-2024 (2024).
4. ECMWF. *IFS Documentation CY43R3 - Part III: Dynamics and numerical procedures*. 3 (ECMWF, 2017).
5. IPCC. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, vol. In Press (Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021).
6. Eyring, V. *et al*. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development* **9**, 1937–1958 (2016).
7. Khairoutdinov, M. F., Blossey, P. N. & Bretherton, C. S. Global system for atmospheric modeling: model description and preliminary results. *Journal of Advances in Modelling Earth Systems* **14**, e2021MS002968, https://doi.org/10.1029/2021MS002968 (2022).
8. Ullrich, P., Reed, K. & Jablonowski, C. Analytical initial conditions and an analysis of baroclinic instability waves in f-and β-plane 3D channel models. *Quarterly Journal of the Royal Meteorological Society* **141**, 2972–2988, https://doi.org/10.1002/qj.2583 (2015).
9. Simmons, A. J. & Hoskins, B. J. The life cycles of some nonlinear baroclinic waves. *Journal of Atmospheric Science* **35**, 414–432, https://doi.org/10.1175/1520-0469 (1978).
10. Beare, R. J. Boundary layer mechanisms in extratropical cyclones. *Quarterly Journal of the Royal Meteorological Society* **133**, 503–515, https://doi.org/10.1002/qj.30 (2007).
11. Carver, G. OpenIFS Horizontal Resolution and Configurations https://confluence.ecmwf.int/display/OIFS/4.3+OpenIFS%3A+Horizontal+Resolution+and+Configurations Accessed: 2025-02-17.
12. ECMWF. *IFS Documentation CY43R3 - Part IV: Physical processes*. 4 (ECMWF, 2017).
13. Bouvier, C., van den Broek, D., Ekblom, M. & Sinclair, V. A.*bouvierc/Baroclinic lifecycles: OpenIFS initial background states and experiments*, https://doi.org/10.5281/zenodo.10592587 (2024).
14. Huntington, D. & Lyrintzis, C. Improvements to and limitations of Latin hypercube sampling. *Probabilistic engineering mechanics* **13**, 245–253, https://doi.org/10.1016/S0266-8920(97)00013-1 (1998).
15. Bouvier, C. Baroclinic-Wave-Simulation-Ensemble-repository https://version.helsinki.fi/dynamic-meteorology-public/Baroclinic-Wave-Simulation-Ensemble-repository Accessed: 2025-04-10.

16. Sparrow, S. *et al*. OpenIFS@ home version 1: a citizen science project for ensemble weather and climate forecasting. *Geoscientific Model Development* **14**, 3473–3486, https://doi.org/10.5194/gmd-14-3473-2021 (2021).
17. Climateprediction.net (CPDN) program. https://eng.ox.ac.uk/climateprediction Accessed: 2024-08-13.
18. Anderson, D. BOINC: a system for public-resource computing and storage. In *Fifth IEEE/ACM International Workshop on Grid Computing*, 4–10, https://doi.org/10.1109/GRID.2004.14 (2004).
19. CSC - IT Center for Science https://csc.fi/en/ Accessed: 2024-08-13.
20. HyperQueue. https://it4innovations.github.io/hyperqueue/v0.19.0/ Accessed: 2024-08-13.
21. Hodges, K. I. A General Method for Tracking Analysis and Its Application to Meteorological Data. *Monthly Weather Review* **122**, 2573 – 2586 (1994).
22. Hodges, K. I. Feature Tracking on the Unit Sphere. *Monthly Weather Review* **123**, 3458 – 3465 (1995).
23. Hodges, K. I. Adaptive Constraints for Feature Tracking. *Monthly Weather Review* **127**, 1362 – 1373 (1999).
24. Bengtsson, L., Hodges, K. I. & Esch, M. Tropical cyclones in a T159 resolution global climate model: comparison with observations and re-analyses. *Tellus A: Dynamic Meteorology and Oceanography*https://doi.org/10.1111/j.1600-0870.2007.00236.x (2007).
25. Bengtsson, L., Hodges, K. I. & Keenlyside, N. Will extratropical storms intensify in a warmer climate? *Journal of Climate* **22**, 2276–2301, https://doi.org/10.1175/2008JCLI2678.1 (2009).
26. Lorenz, E. N. Available potential energy and the maintenance of the general circulation. *Tellus* **7**, 157–167, https://doi.org/10.3402/tellusa.v7i2.8796 (1955).
27. Oort, A. H. On estimates of the atmospheric energy cycle. *Monthly Weather Review* **92**, 483 – 493 (1964).
28. Toropainen, A. Lorenz energy cycle of a baroclinic wave simulation. *Helsingin yliopisto* (2024).
29. Lindzen, R. S. & Farrell, B. A Simple Approximate Result for the Maximum Growth Rate of Baroclinic Instabilities. *Journal of Atmospheric Sciences* **37**, 1648 – 1654 (1980).
30. Attinger, R., Spreitzer, E., Boettcher, M., Wernli, H. & Joos, H. Systematic assessment of the diabatic processes that modify low-level potential vorticity in extratropical cyclones. *Weather and Climate Dynamics* **2**, 1073–1091, https://doi.org/10.5194/wcd-2-1073-2021 (2021).
31. Dolores-Tesillos, E., Teubler, F. & Pfahl, S. Future changes in north atlantic winter cyclones in cesm-le – part 1: Cyclone intensity, potential vorticity anomalies, and horizontal wind speed. *Weather and Climate Dynamics* **3**, 429–448, https://doi.org/10.5194/wcd-3-429-2022 (2022).
32. Schemm, S. & Riviére, G. On the Efficiency of Baroclinic Eddy Growth and How It Reduces the North Pacific Storm-Track Intensity in Midwinter. *Journal of Climate* **32**, 8373 – 8398, https://doi.org/10.1175/JCLI-D-19-0115.1 (2019).
33. Cornér, J., Bouvier, C., Doiteau, B., Pantillon, F. & Sinclair, V. A. Classification of north atlantic and european extratropical cyclones using multiple measures of intensity. *Natural Hazards and Earth System Sciences* **25**, 207–229, https://doi.org/10.5194/nhess-25-207-2025 (2025).
34. Leckebusch, G. C., Renggli, D. & Ulbrich, U. Development and application of an objective storm severity measure for the Northeast Atlantic region. *Meteorologische Zeitung* **17**, 575–587, https://doi.org/10.1127/0941-2948/2008/0323 (2008).
35. Sinclair, V. A. & Catto, J. L. The relationship between extra-tropical cyclone intensity and precipitation in idealised current and future climates. *Weather and Climate Dynamics* **4**, 567–589, https://doi.org/10.5194/wcd-4-567-2023 (2023).
36. Bouvier, C. *et al*. Baroclinic wave simulation ensemble: a machine learning ready dataset. University of Helsinki, Matemaattis-luonnontieteellinen tiedekunta https://doi.org/10.23729/fd-b84c06b2-0950-3bd4-bb98-defc0518eaf5 (2025).
37. Zappa, G., Shaffrey, L. C. & Hodges, K. I. The ability of cmip5 models to simulate north atlantic extratropical cyclones. *Journal of Climate* **26**, 5379 – 5396, https://doi.org/10.1175/JCLI-D-12-00501.1 (2013).
38. Sainsbury, E. M. *et al*. How important are post-tropical cyclones for european windstorm risk? *Geophysical Research Letters* **47**, e2020GL089853, https://doi.org/10.1029/2020GL089853 (2020).
39. Priestley, M. D. K. & Catto, J. L. Future changes in the extratropical storm tracks and cyclone intensity, wind speed, and structure. *Weather and Climate Dynamics* **3**, 337–360, https://doi.org/10.5194/wcd-3-337-2022 (2022).
40. Eady, E. T. Long waves and cyclone waves. *Tellus* **1**, 33–52, https://doi.org/10.1111/j.2153-3490.1949.tb01265.x (1949).
41. Laurila, T. K., Sinclair, V. A. & Gregow, H. Climatology, variability, and trends in near-surface wind speeds over the North Atlantic and Europe during 1979-2018 based on ERA5. *Int. J. Climatol.* **41**, 2253–2278, https://doi.org/10.1002/joc.6957 (2021).
42. CSC. Metax api documentation. https://metax.fairdata.fi/docs/ (2025).
43. CPDN. Contact page. https://climateprediction.net/contact-us/ (2025).

Acknowledgements

The authors wish to acknowledge CSC - IT Center for Science, Finland, for computational resources. The authors want to thank ECMWF for making OpenIFS available to the University of Helsinki. This research was supported by the Research Council of Finland (grant no 338615). JC was partly funded by the University of Helsinki Doctoral School. The authors wish to thank all of the volunteers for giving their computing resources to the CPDN project. Open access funded by Helsinki University Library.

Author contributions

C.B. and V.A.S. designed simulations and thus the dataset. C.B. was responsible for the B.W.S. development, the ensemble definition and processing and the feature extraction. J.C. was responsible for the tracking the ETCs in all of the baroclinic wave simulations. A.B. and G.C. helped with the deployment of the baroclinic wave simulation on the CPDN infrastructure. S.S. and D.W. monitored the computation on CPDN infrastructure and transferred the data. V.A.S. supervised the dataset creation and secured funding. All authors discussed the experiment and results and reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.