Documentation for dataset: Random forest models of magnetospheric ultra-low frequency wave power

S.N.Bentley, J.Stout, T.Bloch, C.E.J.Watt

April 2020

1 Overview of dataset

Title: Random forest models of magnetospheric ultra-low frequency wave power.

Summary: A series of decision tree ensembles (random forests) which have been trained on magnetospheric ultra-low frequency (ULF, 1-15mHz) plasma waves observed by ground based stations. Wave power spectral density is predicted using input parameters of station, frequency, solar wind properties $[v_{sw}, Bz, var(Np)]$ and azimuthal angle (i.e. magnetic local time, or MLT). Constructed using fifteen years of solar wind OMNI data and a latitudinal magnetometer chain from the CANOPUS/CARISMA network. This dataset therefore comprises a statistical model of ULF ground-based wave power throughout the magnetosphere, dependent on solar wind conditions.

Background: Radiation belt models are necessary to describe the near-Earth space environment, which is hostile to spacecraft such as the satellites underpinning modern life. ULF waves are involved in the energisation and transport of radiation belt electrons and are therefore a significant component of such models. Empirical models to predict the wave power enable us to test our understanding of the underlying physics and to predict the resulting radial diffusion of radiation belt electrons. The series of models here are more accurate, easier to use and significantly more versatile than the previous generation.

Publication year: 2020
Creators: S.N.Bentley, J.Stout, T.Bloch, C.E.J.Watt
Organisation: University of Reading
Rights-holders: University of Reading, Jennifer Stout, Téo Bloch
Sources: Solar wind data was obtained from the National Aeronautics and Space Administration/ Goddard Flight Center's OMNI data set via OMNIWeb (`http://omniweb.gsfc.nasa.gov/`, (King & Papitashvili, 2005)) whilst ground-based magnetic field measurements were obtained from the Churchill Line of the CANOPUS magnetometer chain in Canada (Rostoker et al., 1995), now upgraded and expanded into the CARISMA array (Mann et al., 2008).

2 Terms of use

3 Funding

This work was supported by STFC grant ST/R000921/1 and NERC grant NE/P017274/1. T.B.'s PhD is supported by ST/R505031/1. Additional funding from the University of Reading Undergraduate Research Opportunities Programme (UROP) supported J.S. during summer 2019.

4 Construction of dataset

This dataset corresponds to the model presented in *"Random forest models of magnetospheric ultra-low frequency wave power"* (Bentley, Stout, Bloch, & Watt, 2020), where more detail on the data processing and choice of hyperparameters (settings) can be found, along with optimisation and validation of the models.

To summarise, fifteen years of hourly solar wind and ground-based magnetometer observations were used, from 1990-2004 inclusive. Power spectral density (PSD) was calculated for the ground-observed waves using the multitaper method for four stations FCHU, GILL, ISLL and PINA. Log$_{10}(PSD)$ in units of $\frac{(nT)^2}{Hz}$ was the target variable on which our decision trees were trained, with input vectors containing magnetic local time, solar wind speed v_{sw}, north-south component of the interplanetary magnetic field Bz and variance of proton number density $log_{10}(var(Np))$. Ground station magnetic local time $m = [0, 23]$ was converted to be cyclic (i.e. $[m_x = \sin(\frac{2\pi m}{24}), m_y = \cos(\frac{2\pi m}{24})]$). A decision tree ensemble was made for each station, frequency, and horizontal geomagnetic component, reducing mean square error (MSE) and using a minimum number of 18 samples per leaf and a minimum depth of 11. Each ensemble contains 256 trees. Random forests (decision tree ensembles) were trained using Python module `scikit-learn` function `RandomForestRegressor` (Pedregosa et al., 2011).

5 Contents

Each ensemble model is saved in two formats: `joblib` and `json`. Joblib files are a serialisation similar to python `pickle` files, which are quick and easy to use but persistence is not guaranteed between versions. JSON files are human-readable files containing the equivalent model. The Python modules `joblib` (v0.14.1, `https://joblib.readthedocs.io/en/latest/`) and `sklearn-json` (v0.1.0, `https://github.com/mlrequest/sklearn-json`) were used respectively.

File names follow the format `componentSTATION_freqmHz.format`, for example `xGILL_5-0mHz.json`. Possible combinations for these file names are indicated in Table 1.[1] These have been stored in

[1] For completeness we have included all frequencies. However, any spectral estimation unavoidably smooths some of the power across nearby frequencies. In the case of the multitaper method used here, with a time half-bandwidth product of 1.4 applied to five-second resolution hourly data, the resolution bandwidth of the multitaper estimate is seven times the frequency resolution, i.e. $[-W, W] = [-7\Delta f, 7\Delta f] = \pm$ 1.89 mHz. This smearing will drop off with distance. Nevertheless, models trained on neighbouring frequencies may contain some artificially "shared" power; for studies across multiple frequencies, we suggest a resolution of six times

String in filename	Possible options
`component`	`x` (ground geomagnetic north-south), or `y` (ground geomagnetic east-west)
`STATION`	FCHU, GILL, ISLL, PINA
`freq`	Every 0.277 mHz, from 1.67-15 mHz, in format e.g. `1-67, 5-0, 11-67`
`format`	`.joblib` or `.json`

Table 1: All options for filenames

compressed files by `component` and `STATION`, e.g. `xFCHU`.

5.1 Using the dataset

The models can be read in to Python in either format. Example code below (tested in Python (v3.8.1)) can be used to read in and apply models to a `pandas` dataframe of input values:

```
from joblib import dump,load
import sklearn_json as skljson
import pandas as pd
import numpy as np

modeldir = '<model-directory>'
m_str1 = 'xGILL_5-0mHz.joblib'
m_str2 = 'xFCHU_5-0mHz.json'

# load in one of each type of model
model1 = load( modeldir+m_str1 )
model2 = skljson.from_json( modeldir+m_str2 )

# make a minimum example dataframe
mlt = 18
mlt_x = np.sin(2*np.pi*mlt / 24 )
mlt_y = np.cos(2*np.pi*mlt / 24 )

df = pd.DataFrame(
    {'speed':[450,500],
    'Bz':  [0.1,1],
    'var_Np':[-0.7,-0.22], # log10(varNp)
    'MLTx':[mlt_x,mlt_x],
    'MLTy':[mlt_y,mlt_y]} )


# predict log10(PSD) under these conditions for GILL, FCHU 5mHz
Gpow = model1.predict(df)
Fpow = model2.predict(df)
```

More detailed operations can be applied to ensembles and ensemble members directly as per the `scikit-learn` documentation.

the frequency resolution $\Delta f = 0.277$mHz (e.g. 1.67, 3.33 and 5.0 mHz etc.) is a good compromise between well-spaced frequency values with minimal smearing, and remaining physically useful. Alternatively, power spectral density may be integrated across frequencies of interest.

Dependencies: `scitkit-learn` (v0.22.1, (Grisel et al., 2020)), `pandas` (v0.25.3, (Reback et al., 2019)), `numpy` (v1.17.5) and either `joblib` (v0.14.1, `https://joblib.readthedocs.io/`) or `sklearn-json` (v0.1.0, `https://github.com/saromanov/scikit-json`).

References

Bentley, S. N., Stout, J., Bloch, T., & Watt, C. E. J. (2020). Random forest model of ultra-low frequency magnetospheric wave power. *Earth and Space Science*. doi: 10.1029/2020EA001274

Grisel, O., Mueller, A., Lars, Gramfort, A., Louppe, G., Prettenhofer, P., ... Eustache (2020, jan). *scikit-learn/scikit-learn: Scikit-learn 0.22.1.* Retrieved from `https://doi.org/10.5281/zenodo.3596890{#}.Xp8CUik72Mk.mendeley` doi: 10.5281/ZENODO.3596890

King, J. H., & Papitashvili, N. E. (2005). Solar wind spatial scales in and comparisons of hourly Wind and ACE plasma and magnetic field data. *Journal of Geophysical Research: Space Physics*, *110*(A2), 1–9. doi: 10.1029/2004JA010649

Mann, I. R., Milling, D. K., Rae, I. J., Ozeke, L. G., Kale, A., Kale, Z. C., ... Singer, H. J. (2008). *The upgraded CARISMA magnetometer array in the THEMIS era* (Vol. 141) (No. 1-4). doi: 10.1007/s11214-008-9457-6

Pedregosa, F., Varoquax, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved from `http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html`

Reback, J., McKinney, W., den Bossche, J. V., Jbrockmendel, Augspurger, T., Cloud, P., ... Chen, K. W. (2019, oct). *pandas-dev/pandas: v0.25.3.* Retrieved from `https://doi.org/10.5281/zenodo.3524604{#}.Xqv8h{_}DvzU4.mendeley` doi: 10.5281/ZENODO.3524604

Rostoker, G., Samson, J. C., Creutzberg, F., Hughes, T. J., McDiarmid, D. R., McNamara, A. G., ... Cogger, L. L. (1995). Canopus - A ground-based instrument array for remote sensing the high latitude ionosphere during the ISTP/GGS program. *Space Science Reviews*, *71*(1-4), 743–760. doi: 10.1007/BF00751349