A random forest model of ULF wave power

Part I: Summary & Context

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

snbentley@outlook.com

Summary

- Our freely available model predicts magnetospheric ULF wave power (1-15mHz) using solar wind properties.
- We explain why not all physical processes can be extracted from parameterised models, and outline a hypothesis testing framework to iteratively explore driving.
- We use our model to investigate the dawn-dusk power asymmetry. We also find that magnetospheric compression and internal processes should be included in future.

Poster Contents Summary:

Part I: Background

What are ULF waves. Future model uses.

Part II: The Model

We use a machine learning technique (random forests) to predict ULF wave power spectral density (PSD).

Part III: Getting physics from parameterised models

Not all physics can be extracted. We suggest a hypothesis testing technique.

Part IV: Example Physics Results

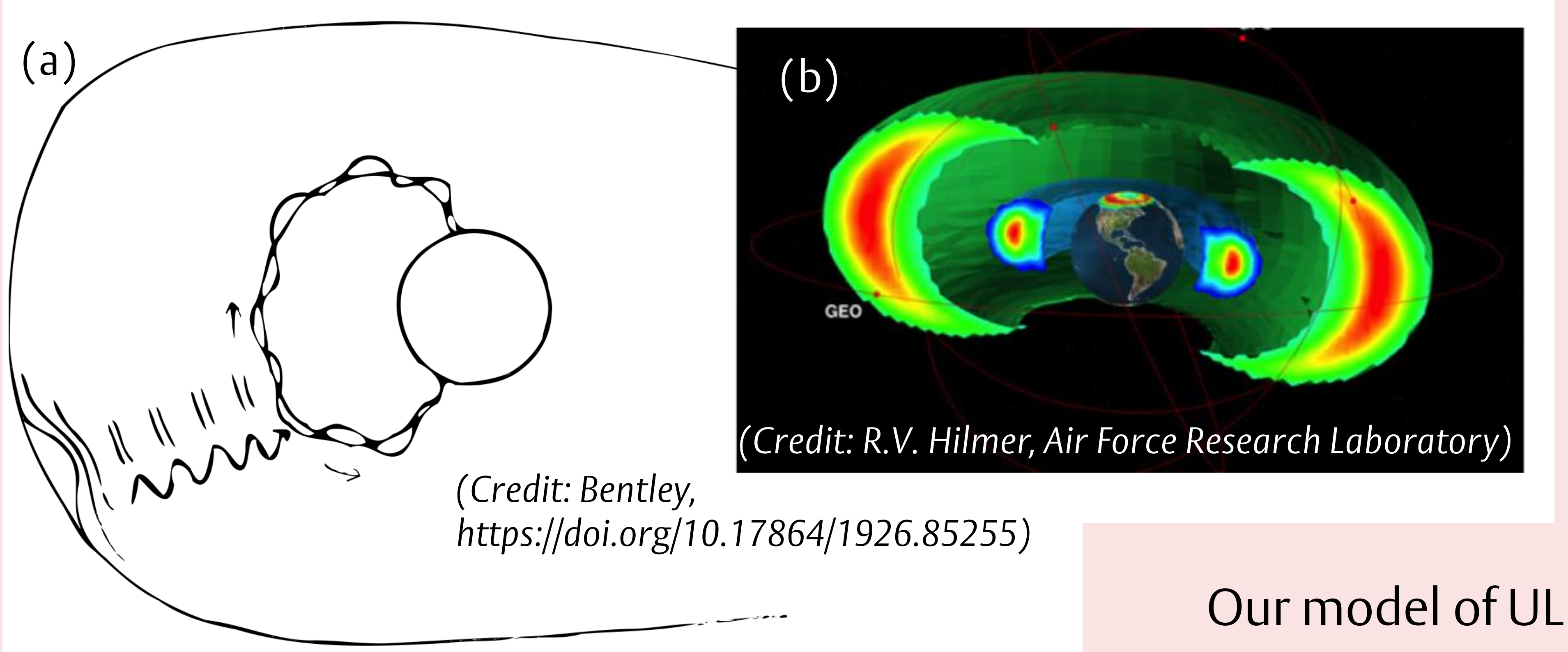
Applying the hypothesis testing method:

MAIN RESULTS:

1. The dawn-dusk wave power asymmetry is a combined effect of radial density profiles and magnetopause perturbations.
2. *var(Np)* does not represent wave driving by magnetopause perturbations.
3. Nor does *Bz*, which likely represents wave power increases with substorms.
4. The internal state of the magnetosphere adds significant uncertainty. It should be included in future.

Context

Ultra-low frequency (ULF) waves are magnetosphere-scale oscillations, with periods of minutes to hours. They affect the energisation and transport of electrons in Earth's radiation belts. (Fig. 1)



(Credit: R.V. Hilmer, Air Force Research Laboratory)

(Credit: Bentley, https://doi.org/10.17864/1926.85255)

Fig. 1 *(a) Perturbations at the magnetopause can drive ULF plasma waves that propagate inwards and couple with the dipole magnetic field. Internal wave sources exist but contribute far less to low frequencies. (b) Earth's radiation belts contain high-energy particles hazardous to spacecraft.*

Our model uses variable bins to mitigate several difficulties inherent to space physics data (**sparseness, interdependent driving parameters, nonlinearity**) to produce an approximation of ULF wave power spectral density depending on the incoming solar wind.

Our model of ULF waves will allow us to

- Investigate global magnetospheric dynamics
- Estimate radial diffusion coefficients, which are a vital component of radiation belt forecasts used to protect spacecraft.
- Investigate azimuthal (magnetic local time, MLT) variation at higher, more appropriate resolution than previously.

References

1. Bentley et al. 2019, *Capturing uncertainty in magnetospheric ultralow frequency wave models*, Space Weather, https://doi.org/10.1029.2018SW002102
2. Bentley et al. 2020, *Random forest model of ultra-low frequency magnetospheric wave power*, Earth and Space Science, https://doi.org/10.1029/2020EA001274
3. Bentley et al. 2020, *Dataset: Random forest model of ultra-low frequency magnetospheric wave power*, https://doi.org/10.5281/zenodo.3828506

Acknowledgements

- I.R. Mann, D.K. Milling and the rest of the CARISMA team for data. CARISMA is operated by the University of Alberta, funded by the Canadian Space Agency.
- NASA/GSFC's Space Physics Data Facilities, omniweb.gsfc.nasa.gov

A random forest model of ULF wave power

Part II: The Model

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

snbentley@outlook.com

What is a random forest?

Decision trees predict continuous variables by iteratively splitting bins along input (driving) parameters based on mean square error (MSE) in the value being predicted.

Example: if power spectral density (PSD) is greater with increasing solar wind speed, an initial decision may be to create two bins split at $v_{sw} \sim 450$ km s^{-1}.

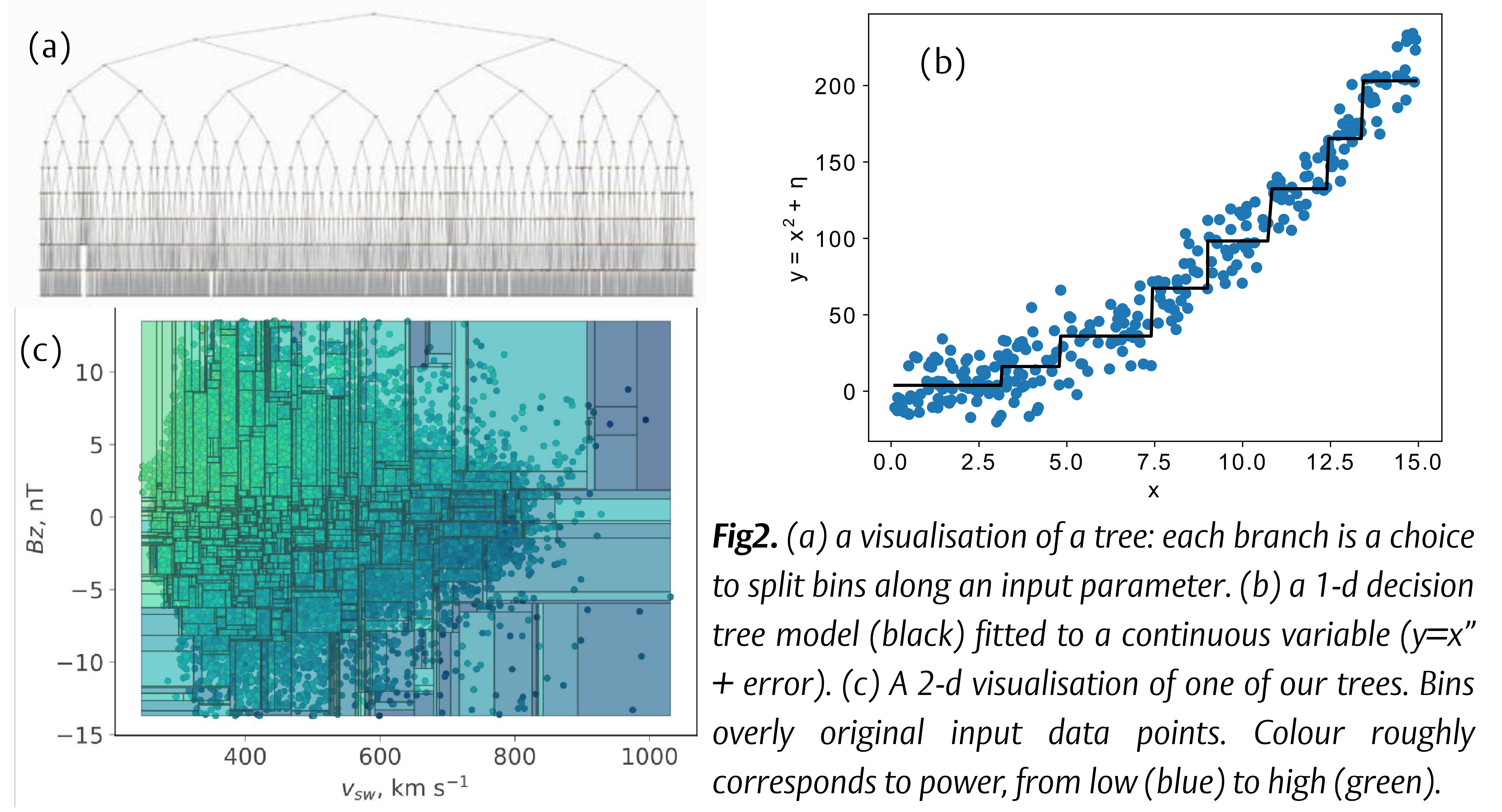


Fig2. *(a) a visualisation of a tree: each branch is a choice to split bins along an input parameter. (b) a 1-d decision tree model (black) fitted to a continuous variable (y=x" + error). (c) A 2-d visualisation of one of our trees. Bins overly original input data points. Colour roughly corresponds to power, from low (blue) to high (green).*

Random forests (decision tree ensembles) are used to reduce variance and generalise well to new data. Each ensemble contains 256 trees trained on different subsets of the data (drawn with replacement). The final predicted value is the mean output of all 256 trees.

Benefits:

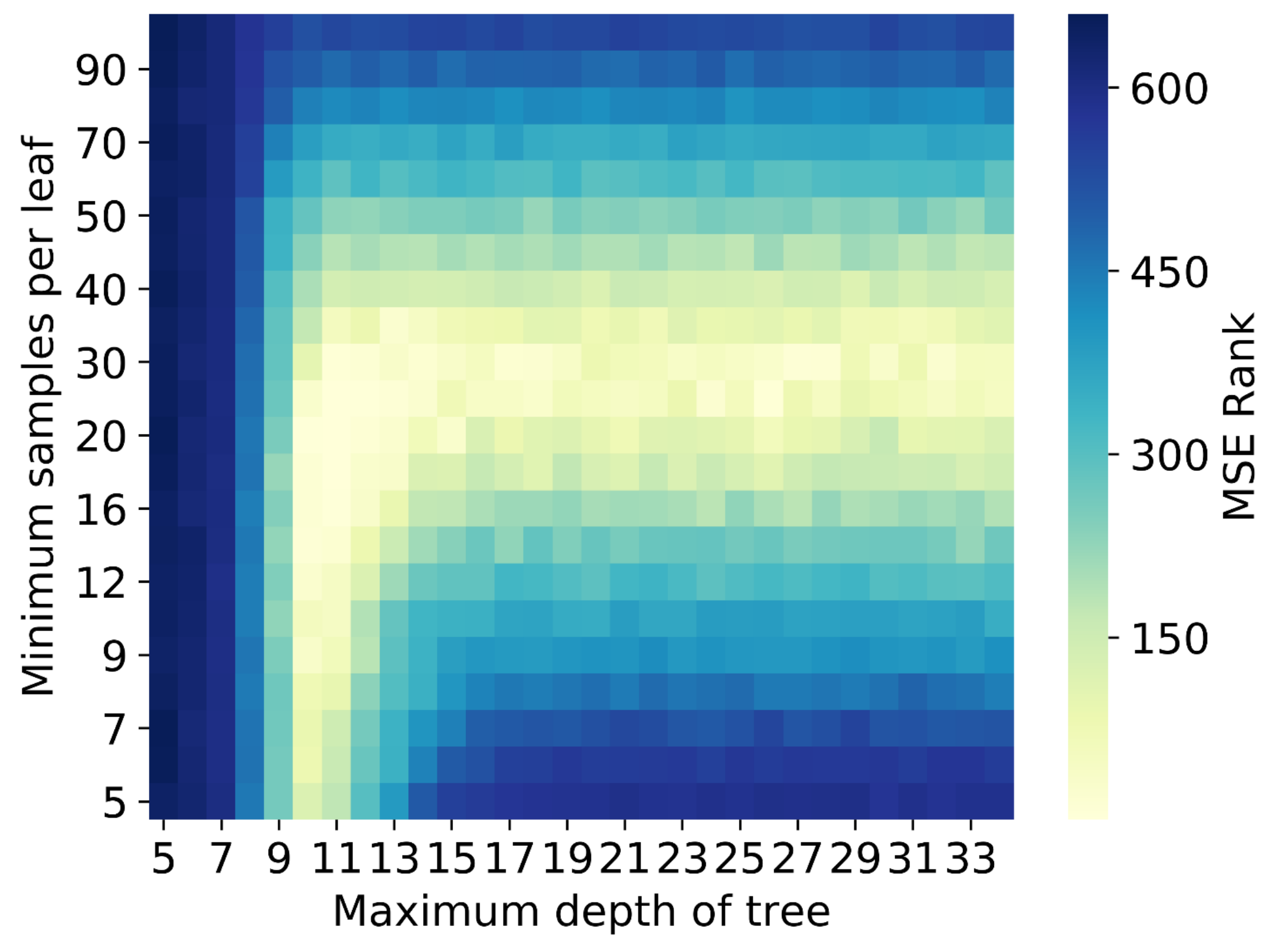
- Variable bins mean large bins for sparse data points, they also capture rapid changes in power, e.g. at *Bz=0* , Fig 2(c)
- No assumption that power is linear along input parameters; power is represented using a series of step functions, Fig 2(b).

Our model

We predict ULF wave PSD across 1-15mHz, for each horizontal component, using four CARISMA ground stations at different latitudes (*L*~4.21 to 7.94), using magnetic local time and solar wind speed, *Bz* and variance of proton number density from OMNI Web.

$$[\boldsymbol{freq}, \boldsymbol{latitude}, \boldsymbol{comp.}, \boldsymbol{MLT}, \boldsymbol{vsw}, \boldsymbol{Bz}, \boldsymbol{var}(\boldsymbol{Np})] \rightarrow \boldsymbol{PSD}$$

Fig 3. *Many model settings were tested and ranked by their mean square error (MSE, calculated using 5-fold cross-validation). We use min_samples_per_leaf = 18 and maximum_depth = 11 to minimise the MSE.*



Model testing

Before training final models on the full 1990-2005 dataset, we estimate mean square error using subsets. Largest and smallest median MSE were 0.68 and 0.13 $\log_{10}$(PSD) (nT)^2/Hz respectively.

We compare different models using forecasting skill. (Table 1) Decision tree ensembles perform best.

Table 1 *Positive skill scores indicate that the tested model is better than a reference model (here, a "random" model sampling the original power distribution). Decision tree ensembles outperform the previous model and using time lags of either 24h or 1h. We show results for GILL 3.33 mHz.*

$$Skill = 100\left(1 - \frac{MSE_1}{MSE_2}\right)$$

Model tested	Skill
Random forest	81.2
Previous model[2]	78.0
24 h lag	37.4
1 h lag	73.9

The model is available via Zenodo, including full documentation and usage examples.

A random forest model of ULF wave power

PART III: GETTING PHYSICS FROM PARAMETERISED MODELS

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

snbentley@outlook.com

A parameterised model

Our random forests are parameterised models that approximate the average value of power as a surface, (a single value in our chosen parameter space) with all contributing processes compounded together.

$$[\boldsymbol{freq}, \boldsymbol{latitude}, \boldsymbol{comp.}, \boldsymbol{MLT}, \boldsymbol{vsw}, \boldsymbol{Bz}, \boldsymbol{var}(\boldsymbol{Np})] \rightarrow \boldsymbol{PSD}$$

A model that successfully represents the physical output is not necessarily well suited to investigating individual contributing processes.

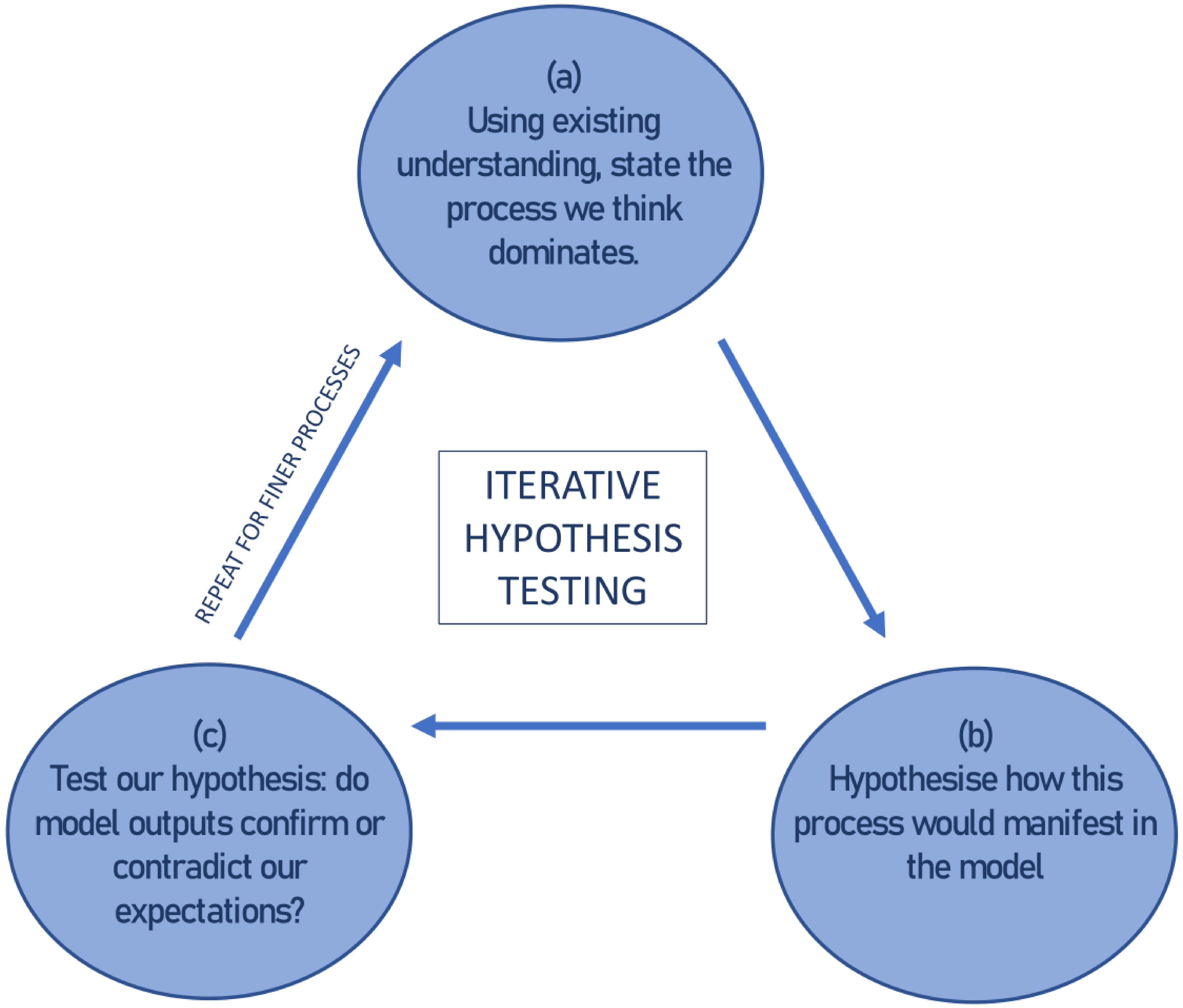
Difficulties extracting physics

Assuming one can extract all underlying physical processes is equivalent to assuming each process adds linearly to the final PSD. In fact our approximation is **strongly nonlinear:** driving parameters depend on each other, and driving processes are related to multiple parameters and processes.

- **May only be able to extract dominant driving processes:** lesser processes may be indistinguishable in the final, convolved PSD.
  Solution: Iterative hypothesis testing.
- **Choice of parameters:** To disentangle these processes in analysis, parameter axes must directly relate to PSD.
  *Solution***:** Use three dominant solar wind parameters causally correlated to ULF PSD.
- **Uncertainty is useful:** it tells us where in parameter space represents the physics is less well represented.
  Solution: Include hypotheses about uncertainty.
- **Bias and interdependence** often eclipse individual processes. For example, our model inherently includes pre-Earth solar wind processing.
  Solution: Cannot be fully solved. Mitigated by parameter choice and variable bins.

Physics via successive hypothesis testing

We suggest a hypothesis testing framework to examine the physics driving ULF wave power. This formalises the approach taken in full statistical surveys, beginning with dominant driving processes, testing how they manifest in the model, and then examining remaining power.



(a)
Using existing understanding, state the process we think dominates.

(b)
Hypothesise how this process would manifest in the model

(c)
Test our hypothesis: do model outputs confirm or contradict our expectations?

ITERATIVE HYPOTHESIS TESTING

Fig 4. *Our suggested hypothesis testing framework.*

Hypothesis testing #1: MLT variation

We test whether the well-documented dawn-dusk wave power asymmetry is due to the combined effect of magnetopause perturbations and different radial plasma density profiles.

> ***Example Hypothesis 1**: Waves driven by external magnetopause perturbations, particularly large amplitude ones, will have more power in the dawn sector. This asymmetry will be larger at higher latitudes.*

In summary, if we see this signature for one (but not all) of the driving parameters which correspond to magnetopause perturbations, we can accept the hypothesis that this is due to the combined effect listed above. We see this signature for v_{sw} (corresponding to Kelvin-Helmholtz instabilities, Fig. 5) but not for *Bz* or *var(Np)*. *(Full results & logic available in paper)*

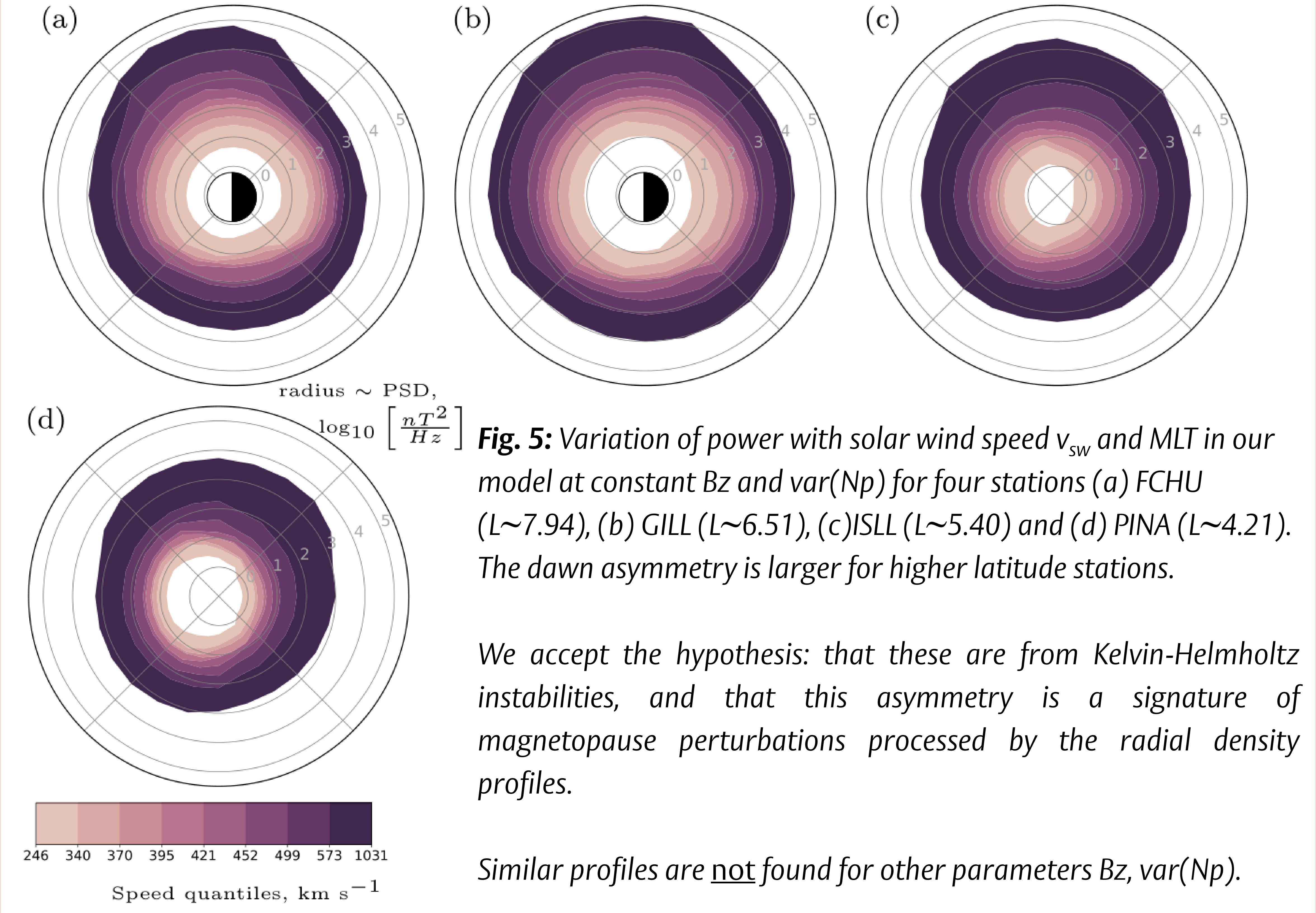


Fig. 5: *Variation of power with solar wind speed v_{sw} and MLT in our model at constant Bz and var(Np) for four stations (a) FCHU (L~7.94), (b) GILL (L~6.51), (c)ISLL (L~5.40) and (d) PINA (L~4.21). The dawn asymmetry is larger for higher latitude stations.*

We accept the hypothesis: that these are from Kelvin-Helmholtz instabilities, and that this asymmetry is a signature of magnetopause perturbations processed by the radial density profiles.

Similar profiles are not found for other parameters Bz, var(Np).

Dawn-dusk asymmetry results summary:

- The dawn-dusk wave power asymmetry is a combined effect of the different radial density profiles and wave driving from magnetopause ("external") perturbations such as Kelvin-Helmholtz instabilities.
- We cannot account for the effects of a compressed magnetosphere, but *var(Np)* does not represent wave driving by magnetopause perturbations.
- Nor does *Bz*, which instead likely represents wave power increases with substorms.

Part IV: Example Physics Results

Hypothesis testing #2: Uncertainty

We also test our assumptions about what additional physics we believe needs to be included.

> ***Example Hypothesis 2**: As we do not include substorms in our parameterisation, the greatest remaining uncertainty will be in the midnight sector and uncertainties will be larger for Bz < 0 than Bz > 0.*

In fact we find that other physics may be missing instead. (Fig. 6)

Uncertainty results summary:

- *Bz>0* has greater uncertainty. This is probably because there is larger variability in substorm occurrence for *Bz>0* than *Bz<0*, so *Bz<0* is an better hourly substorm proxy.
- Greatest remaining uncertainty is found for *Bz>0*, low v_{sw} and low *var(Np)*, i.e. when there is the least solar wind driving. This suggests that the configuration of the magnetosphere and internal processes are secondary effects that should be included in future.

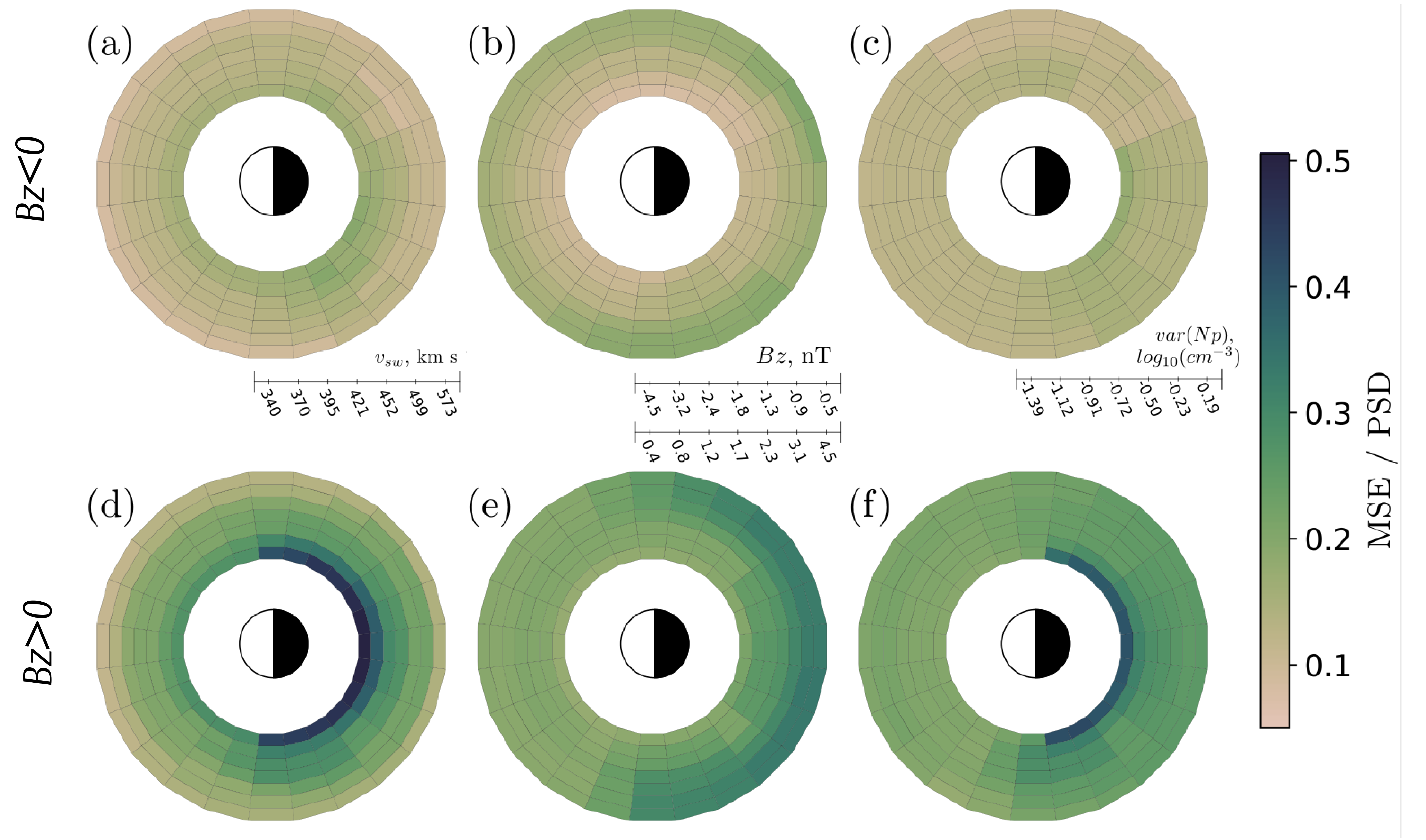


Fig. 6: *Remaining uncertainty by parameter and MLT for constant, median speed, Bz and var(Np) values (421 km s^{-1} , -1.8 and 1.7 nT, and -0.716 log 10 (cm^{-3}) respectively. (a) Uncertainty at speed quantiles for each MLT (quantile values shown in the corner bar) at constant Bz and var(Np). (b) uncertainty for Bz < 0 at median values of speed and var(Np), (c) is uncertainty for var(Np) by MLT. (d)-(f) show the same for Bz > 0.*