0.1 Datasets

0.1.1 Introduction

To accurately predict solar wind parameters, it is imperative to use datasets that contain precise measurements and a cadence high enough to capture both short and long-term trends. In this study, we make use of the OMNI COHO dataset with hourly cadence (**?**). This dataset is formulated in the radial-tangential-normal (RTN) coordinate system, meaning spatial correlations which would otherwise would be present are eliminated, leaving only temporal correlations (**?**). In this study, we focus on radial magnetic field strength (B_r) and radial wind velocity (v_r).

0.1.2 Solar cycles

In order to ensure our training, validation and testing datasets contain similar information for predictions to be accurate, we slice the data by solar cycle. Cycles 21 - 22 are used in the training dataset, cycle 23 is used for validation, and cycle 24 is used for testing. Table 0.1.2 shows the number of points in each cycle, the date ranges of these cycles, and how many NaN points are in each.

Cycle	Start	End	Total Points	% NaN (B_r)	% NaN (v_r)
21	1976-03-01	1986-09-01	92073	35.91	33.94
22	1986-09-01	1996-08-01	86937	49.19	50.21
23	1996-08-01	2008-12-01	108131	0.20	0.35
24	2008-12-01	2019-12-01	96418	0.11	0.15

Table 0.1: Percentage of points that are NaN in each solar cycle.

0.2 Analogue Ensemble

We employ the use of an analogue ensemble (AnEn) **?**, **?** as a baseline metric to measure our LSTM against. Given a recent observation of n points, the AnEn forecasts the next m points. It involves looking at previous observations to find similar ones to the recent observation. A number of analogues k is selected with a certain loss metric - we use mean squared error (MSE) loss to find the analogues that best fit the recent observation.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i = \hat{Y}_i)^2 \tag{0.1}$$

The time for a single prediction depends on the size of the dataset the AnEn can use for analogues.