
Scalable Machine Learning Analysis of Parker Solar Probe Solar Wind Data

Daniela Martin
University of Delaware
dmartin@udel.edu

Connor O'Brien
Boston University
obrienco@bu.edu

Valmir P. Moraes Filho
Catholic University of America
moraesfilho@cua.edu

Jinsu Hong
Georgia State University
jhong36@gsu.edu

Jasmine R. Kobayashi
Southwest Research Institute
jasmine.kobayashi@swri.org

Evangelia Samara
NASA Goddard Space Flight Center
evangelia.samara@nasa.gov

Joseph Gallego
Drexel University
jg3959@drexel.edu

Abstract

We present a scalable machine learning framework for analyzing Parker Solar Probe (PSP) solar wind data using distributed processing and the quantum-inspired Kernel Density Matrices (KDM) method. The PSP dataset (2018–2024) exceeds 150 GB, challenging conventional analysis approaches. Our framework leverages Dask for large-scale statistical computations and KDM to estimate univariate and bivariate distributions of key solar wind parameters, including solar wind speed, proton density, and proton thermal speed, as well as anomaly thresholds for each parameter. We reveal characteristic trends in the inner heliosphere, including increasing solar wind speed with distance from the Sun, decreasing proton density, and the inverse relationship between speed and density. Solar wind structures play a critical role in enhancing and mediating extreme space weather phenomena and can trigger geomagnetic storms; our analyses provide quantitative insights into these processes. This approach offers a tractable, interpretable, and distributed methodology for exploring complex physical datasets and facilitates reproducible analysis of large-scale in situ measurements. Processed data products and analysis tools are made publicly available to advance future studies of solar wind dynamics and space weather forecasting.

1 Introduction

The Parker Solar Probe (PSP), launched in 2018, aims to unveil the mechanisms driving the heating and acceleration of the solar wind [18]. This work focuses on analyzing PSP measurements of solar wind plasma properties obtained by the SWEAP instrument [6]. The dataset spanning 2018 – 2024 exceeds 150 GB, presenting challenges for scalable and interpretable analysis of high-volume time series data. Prior studies seeking to determine the overall properties of the solar wind from data taken in-situ have been based on smaller datasets with poorer temporal resolution in more limited spatial regimes [14, 23], and focus on basic statistical techniques such as binning and averaging.

Traditional approaches, including feature extraction, probabilistic models, and deep learning methods, face limitations when applied to solar wind measurements at this scale [10, 11]. For instance, kernel density estimation (KDE) [16] becomes computationally prohibitive, while normalizing flows

[19, 7] and autogressive models are often difficult to train. Variational autoencoders only approximate the underlying distribution, and implicit generative models such as Generative Adversarial networks [9] or diffusion models [2] do not provide an explicit density function. To address these challenges, we develop a distributed framework using Dask [3] and apply the quantum-inspired Kernel Density Matrices (KDM) method [8] for scalable density estimation and anomaly detection. The computation of the statistical quantities are parallelized, while not a major contribution for Dask itself, represents a practical application for the heliophysics community.

Our contributions are threefold: (1) statistical characterization of solar wind properties using distributed processing; (2) application of KDM to uncover multi-parameter distributions and relationships in large PSP solar wind datasets; and (3) open-source tools for analyzing large in situ datasets.

2 Background

González et al. [8] define KDM over a set \mathbb{X} as a triplet $\rho = (\mathbf{C}, \mathbf{p}, \theta)$ where ρ represents a density matrix defined by

$$f_\rho(\mathbf{x}) = \sum_{\mathbf{x}^{(i)} \in \mathbf{C}} p_i k_\theta^2(\mathbf{x}, \mathbf{x}^{(i)}),$$

Here, \mathbf{C} is the set of components, \mathbf{x} are the original points, and $\mathbf{x}^{(i)}$ is the i -th component of \mathbf{C} within \mathbb{X} . Each component has an associated mixture weight $p_i \in \mathbb{R}$, representing its probability. The kernel function $k_\theta : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ is defined such that $k(\mathbf{x}, \mathbf{x}) = 1$. Using a Gaussian kernel, each component $\mathbf{x}^{(i)}$ is assumed to follow a Gaussian distribution with a normalization constant. Components define a categorical distribution, where each category corresponds to a normal distribution with mean equal to the component value and standard deviation $\sigma/2$, i.e. $\mathcal{N}(x^{(i)}, \sigma/2)$.

The parameters $(\mathbf{C}, \mathbf{p}, \theta)$ are learned by maximizing the log-likelihood of the observed data:

$$\max_{\mathbf{C}, \mathbf{p}, \theta} \sum_{i=1}^{\ell} \log \hat{f}_\rho(x_i),$$

where \hat{f}_ρ is the projection function associated with the KDM density matrix ρ of the component \mathbf{x}_i . The model is trained using automatic differentiation. This approach allows us to compute both univariate and joint distributions, enabling subsequent statistical analyses and anomaly detection. The model can also be used to sample new points from the learned distribution.

3 Experimental Setup

Dataset. The Parker Solar Probe (PSP) is a mission launched on Aug. 12, 2018, by NASA, which has completed 24 orbits as of Jun. 19, 2025. Its primary goal is to better understand the structure and dynamics of the inner corona of the Sun, where the solar wind is heated and accelerated. Determining the processes that cause the acceleration of energetic particles is of great importance to the heliophysics community [18]. PSP carries four instruments: FIELDS (measuring electric and magnetic fields), SWEAP (sampling charged particles such as electrons, protons, and alpha particles), WISPR (imaging the solar corona), and IS \odot IS (sampling high-energy particles) [6]. In this work, we focus on SWEAP data to analyze the plasma properties of the solar wind [13, 22, 1].

The SWEAP data must be preprocessed before it can be used for this study. The SWEAP data are first downloaded as CDF files and converted to Zarr format [4] to enable distributed computation. Fill values due to data outages are flagged and removed, and PSP's radial distance to the Sun is computed and stored alongside the plasma data. The plasma parameters of interest are derived from the raw instrument data either by fitting several Gaussians to the particle distributions measured by the instruments or by taking moments of the particle distributions (see [13, 22, 1] for details). Here we use data derived using the fit technique due to their improved temporal coverage.

Processing Architecture. Our processing architecture comprises two parts: (1) distributed processing using Dask and (2) KDM for estimating univariate and joint distributions of PSP solar wind parameters. The PSP data are processed using Dask Array across multiple cores and nodes. The analysis was performed in two stages: first, Dask was used to compute statistical metrics from the PSP solar wind dataset; second, KDM was applied to generate univariate and bivariate distributions

of parameters such as solar wind speed, proton density, and proton thermal speed. These parameters were selected because they represent the primary physical quantities governing solar wind behavior and provide complementary information on both its bulk motion and thermal state, making them ideal for studying correlations and variability in the inner heliosphere. All analyses were conducted as a function of heliocentric distance, ranging from 0 to 1 astronomical units (AU) in increments of 0.1 AU.

Codebase. Our codebase will be released on GitHub after the anonymized review process.

Hardware. All experiments were conducted on Google Cloud Platform using a c2-standard-8 VM instance (8 vCPUs, 32 GB RAM) with 1 TB SSD storage and a Nvidia L4 (24 GB V-GPU).

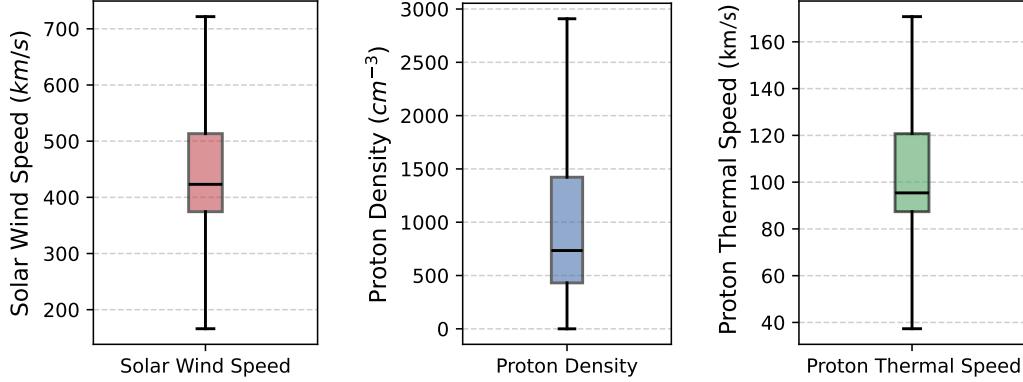


Figure 1: Boxplots of three PSP solar wind parameters from 2018 to 2024: (left) solar wind speed [km/s], (middle) proton density [cm⁻³], and (right) proton thermal speed [eV].

4 Results and Discussion

Several statistics computed with Dask reveal notable features in the PSP solar wind data. Figure 1 shows boxplots for solar wind speed, proton density, and proton thermal speed for the entire dataset (2018 to 2024). Extreme solar wind speed values range from 100 km/s to 1980 km/s, likely caused by instrument artifacts, with a median of 423 km/s. Proton density spans 0.00657–2,560 cm⁻³, from a possible “disappearing solar wind” event [5] at the lower bound to either a genuine coronal feature or an outlier at the upper bound. Its median value of 734 cm⁻³, agrees with previous PSP observations [15]. The proton thermal speed (wp_fit) ranges from 4 km/s (0.084 eV, a potential outlier, to 109,900 km/s (63 MeV), values that almost certainly reflect instrument error. Its median of 95.4 km/s (46 eV) aligns with typical coronal and inner heliosphere conditions [14].

Table 1 summarizes the KDM hyperparameters chosen for all experiments, including the number of components, Gaussian kernel width (σ), and learning rate.

Table 1: KDM hyperparameters. The number of components, selected empirically, provides a balance between model flexibility and physical interpretability. Fewer components tend to underfit the data, yielding smoother density estimates with slightly higher mean error but improved generalization, whereas an excessive number produces spikier densities that tend to overfit the data.

Hyperparameter	Value
Number of Components	400, 800, 1600
σ	0.1 (trainable)
Learning Rate	10^{-3}

Figure 2 shows the univariate distributions of PSP solar wind speed from 0.1 to 0.6 AU in increments of 0.1 AU. Although the analyses covered the full range of 0 to 1 AU, only the subset up to 0.6 AU is shown, since beyond this point the distributions become sparse and noisy due to fewer PSP measurements, which produced spiky curves that are not statistically robust. Our focus on 0.1 to

0.6 AU highlights the near-Sun region where data density is highest and the physical insights are therefore more reliable. The cumulative distribution function (CDF) indicates that solar wind speeds are lower closer to the Sun, consistent with coronal measurements before acceleration [17]. Between 0.2 (orange) and 0.4 AU (red), speeds appear relatively uniform, though at 0.3 AU (green) both fast and slow solar streams emerge.

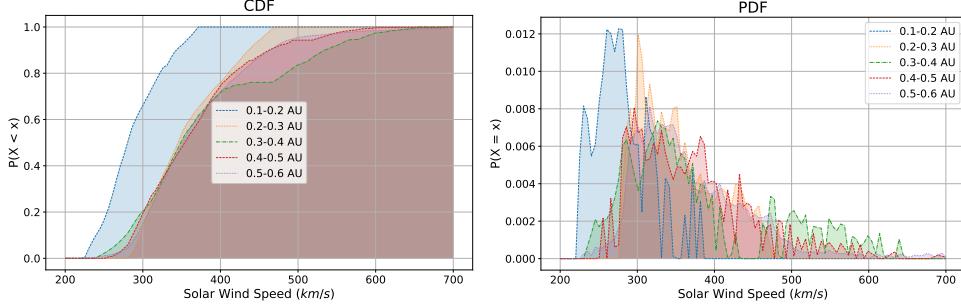


Figure 2: Univariate distributions of PSP’s solar wind speed. **Left:** Cumulative Distribution Function (CDF). **Right:** Probability Density Function (PDF). Each curve corresponds to a different heliocentric distance (0–0.6 AU), as indicated in the legend. Distances beyond 0.6 AU are omitted due to sparse measurements and increased noise.

The probability density function (PDF) in Figure 2 exhibits a right-skewed distribution at all distances. At 0.2 AU (orange), the distribution is leptokurtic, reflecting relatively homogeneous velocities, whereas at 0.3 AU (green) it becomes platykurtic, indicating broader variations. The presence of a large high-speed tail at 0.3 AU, particularly above 450 km/s, suggests increased variability in the solar wind speed, likely associated with transient events or dynamic coronal structures. Together, these trends provide insights into the acceleration and variability of the solar wind in the inner heliosphere.

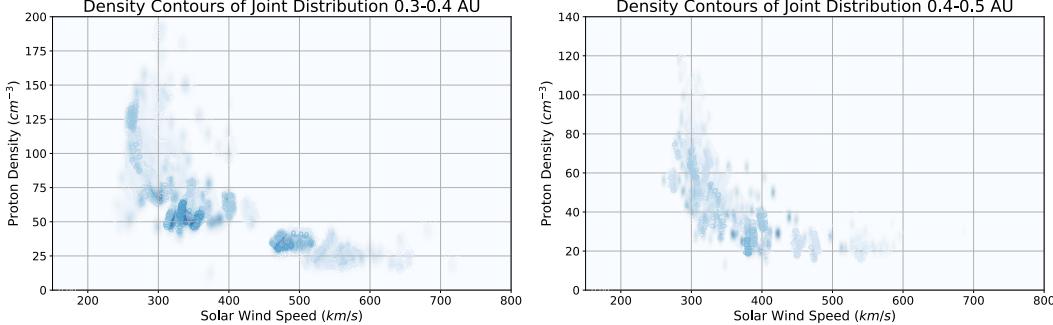


Figure 3: Bivariate distributions of PSP’s solar wind speed versus proton density at different heliocentric distances (AU). **Left panel:** 0.3–0.4 AU. **Right panel:** 0.4–0.5 AU.

Figure 3 illustrates the bivariate distribution of solar wind speed versus proton density at different heliocentric distances. The left panel shows the distribution between 0.3 and 0.4 AU, revealing a clear hyperbolic inverse relationship consistent with a “textbook” understanding of solar wind structures [12]: as solar wind speed increases, proton density decreases. Notably, there is considerable variability in the proton density when the solar wind speed is between 250 and 350 km/s, which may indicate multiple sources or types of solar wind within this slower population. The right panel presents the same relationship for 0.4 to 0.5 AU, where the hyperbolic inverse trend remains. The primary difference is a reduction in the maximum observed solar wind speed, which decreases from approximately 700 km/s at 0.3–0.4 AU range to around 600 km/s at 0.4–0.5 AU, along with a decrease in the maximum proton density from 200 cm^{-3} to 120 cm^{-3} . The density reduction is consistent with rarefaction of the solar wind as it expands radially [20, 21].

We focus on the univariate distributions of solar wind speed and the bivariate distributions of solar wind speed versus proton density in the main text because these parameters are the most represen-

tative of solar wind dynamics. The solar wind speed captures the bulk flow behavior, while its relationship with proton density illustrates key physical phenomena such as the inverse correlation observed in the inner heliosphere. Other parameters and combinations are reported in the supplemental material. The supplemental material provides the corresponding univariate and bivariate distributions for all other solar wind parameters and combinations.

5 Conclusion

We present a scalable and interpretable framework for analyzing Parker Solar Probe solar wind measurements, combining distributed computation with the quantum-inspired KDM method. By computing univariate and bivariate distributions across heliocentric distances, our approach captures key physical relationships, including the inverse correlation between solar wind speed and proton density, as well as variability in kurtosis and skewness reflecting plasma dynamics in the inner heliosphere. Limiting the main analysis to solar wind speed and proton density allows us to focus on the most physically relevant quantities, while additional distributions and parameter combinations are provided in the supplemental material.

Our methodology is efficient, adaptable to other high-volume space physics datasets, and facilitates reproducible analysis of large-scale in situ measurements. The framework also enables the identification of anomalies and extreme events, highlighting its potential for uncovering physically meaningful patterns and supporting future space weather studies.

Broader Impact

Understanding solar wind phenomena is of paramount importance for deciphering the fundamental drivers of extreme solar events such as coronal mass ejections. Moreover, comparing observations from the Parker Solar Probe (PSP) with established physical models enhances confidence in forecasting frameworks developed using PSP data.

Acknowledgements

This work is a research product of Heliolab (heliolab.ai), an initiative of the Frontier Development Lab (FDL.ai). FDL is a publicprivate partnership between NASA, Trillium Technologies (trillium.tech), and commercial AI partners including Google Cloud and NVIDIA. Heliolab was designed, delivered, and managed by Trillium Technologies Inc., a research and development company focused on intelligent systems and collaborative communities for Heliophysics, planetary stewardship and space exploration. We gratefully acknowledge Google Cloud for extensive computational resources enabled through VMware. This material is based upon work supported by NASA under award No. 80GSFC23CA040. Any opinions, findings, and conclusions or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the National Aeronautics and Space Administration.

References

- [1] Case , A. W., Kasper , J. C., Stevens , M. L., Korreck , K. E., Paulson , K., Daigneau , P., Caldwell , D., Freeman , M., Henry , T., Klingensmith , B., Bookbinder , J. A., Robinson , M., Berg , P., Tiu , C., Wright , K. H., Reinhart , M. J., Curtis , D., Ludlam , M., Larson , D., Whittlesey , P., Livi , R., Klein , K. G., & Martinovi , M. M. (2020) The Solar Probe Cup on the Parker Solar Probe. *The Astrophysical Journal Supplement Series* **246**(2):43.
- [2] Croitoru , F.-A., Hondru , V., Ionescu , R. T., & Shah , M. (2023) Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(9):10850–10869.
- [3] Dask Development Team . 2016.
- [4] Developers , Z. zarr-python: Chunked, compressed, n-dimensional arrays for python (version 3.1.2). 2025. Accessed: 2025-08-27.
- [5] Fowler , C. M., Shaver , S., Hanley , K. G., Andersson , L., McFadden , J., Mitchell , D., Halekas , J., Ma , Y., Espley , J., & Curry , S. (2024) Disappearing Solar Wind at Mars: Changes in the MarsSolar Wind Interaction. *Journal of Geophysical Research: Space Physics* **129**(1):e2023JA031910.

- [6] Fox , N. J., Velli , M. C., Bale , S. D., Decker , R., Driesman , A., Howard , R. A., Kasper , J. C., Kinnison , J., Kusterer , M., Lario , D., Lockwood , M. K., McComas , D. J., Raouafi , N. E., & Szabo , A. (2016) The Solar Probe Plus Mission: Humanity First Visit to Our Star. *Space Science Reviews* **204**(1-4):7–48.
- [7] Gallego-Mejia , J. A. & González , F. A. (2023) Demande: Density matrix neural density estimation. *IEEE Access* **11**:53062–53078.
- [8] González , F. A., Ramos-Pollán , R., & Gallego-Mejia , J. A. (2025) Kernel density matrices for probabilistic deep learning. *Quantum Machine Intelligence* **7**(94).
- [9] Goodfellow , I. J., Pouget-Abadie , J., Mirza , M., Xu , B., Warde-Farley , D., Ozair , S., Courville , A., & Bengio , Y. (2014) Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, (eds.), *Advances in Neural Information Processing Systems 27*, . Curran Associates, Inc.
- [10] Henderson , T. & Fulcher , B. D. (2021) An empirical evaluation of time-series feature sets. In *2021 International Conference on Data Mining Workshops (ICDMW)* pages 1032–1038. IEEE.
- [11] Jadon , S., Milczek , J. K., & Patankar , A. (2021) Challenges and approaches to time-series forecasting in data center telemetry: A survey. *arXiv preprint arXiv:2101.04224*
- [12] Kallenrode , M.-B. (2010) *Space physics: an introduction to plasmas and particles in the heliosphere and magnetospheres ; with 12 tables, numerous excercises and problems*. Berlin Heidelberg: Springer, 3. ed., paperback ed edition.
- [13] Kasper , J. C., Abiad , R., Austin , G., Balat-Pichelin , M., Bale , S. D., Belcher , J. W., Berg , P., Bergner , H., Berthomier , M., Bookbinder , J., Brodu , E., Caldwell , D., Case , A. W., Chandran , B. D. G., Cheimets , P., Cirtain , J. W., Cranmer , S. R., Curtis , D. W., Daigneau , P., Dalton , G., Dasgupta , B., DeTomaso , D., Diaz-Aguado , M., Djordjevic , B., Donaskowski , B., Effinger , M., Florinski , V., Fox , N., Freeman , M., Gallagher , D., Gary , S. P., Gauron , T., Gates , R., Goldstein , M., Golub , L., Gordon , D. A., Gurnee , R., Guth , G., Halekas , J., Hatch , K., Heerikuisen , J., Ho , G., Hu , Q., Johnson , G., Jordan , S. P., Korreck , K. E., Larson , D., Lazarus , A. J., Li , G., Livi , R., Ludlam , M., Maksimovic , M., McFadden , J. P., Marchant , W., Maruca , B. A., McComas , D. J., Messina , L., Mercer , T., Park , S., Peddie , A. M., Pogorelov , N., Reinhart , M. J., Richardson , J. D., Robinson , M., Rosen , I., Skoug , R. M., Slagle , A., Steinberg , J. T., Stevens , M. L., Szabo , A., Taylor , E. R., Tiu , C., Turin , P., Velli , M., Webb , G., Whittlesey , P., Wright , K., Wu , S. T., & Zank , G. (2016) Solar Wind Electrons Alphas and Protons (SWEPAP) Investigation: Design of the Solar Wind and Coronal Plasma Instrument Suite for Solar Probe Plus. *Space Science Reviews* **204**(1-4):131–186.
- [14] Ma , X., Nykyri , K., Dimmock , A., & Chu , C. (2020) Statistical Study of Solar Wind, Magnetosheath, and Magnetotail Plasma and Field Properties: 12+ Years of THEMIS Observations and MHD Simulations. *Journal of Geophysical Research: Space Physics* **125**(10):e2020JA028209.
- [15] Maruca , B. A., Qudsi , R. A., Alterman , B., Walsh , B. M., Korreck , K. E., Verscharen , D., Bandyopadhyay , R., Chhiber , R., Chasapis , A., Parashar , T. N., & others (2023) The trans-heliospheric survey-radial trends in plasma parameters across the heliosphere. *Astronomy & Astrophysics* **675**:A196.
- [16] Parzen , E. (1962) On estimation of a probability density function and mode. *The annals of mathematical statistics* **33**(3):1065–1076.
- [17] Patel , R., Niembro , T., Xie , X., Seaton , D. B., Badman , S. T., Roy , S., Rivera , Y. J., Reeves , K. K., Stenborg , G., Hess , P., West , M. J., Feller , A., Hirzberger , J., Orozco Suárez , D., Solanki , S. K., Strecker , H., & Valori , G. (2025) Direct in situ observations of eruption-associated magnetic reconnection in the solar corona. *Nature Astronomy*
- [18] Raouafi , N. E., Matteini , L., Squire , J., Badman , S., Velli , M., Klein , K., Chen , C., Matthaeus , W., Szabo , A., Linton , M., & others (2023) Parker solar probe: Four years of discoveries at solar cycle minimum. *Space Science Reviews* **219**(1):8.
- [19] Rezende , D. & Mohamed , S. (2015) Variational inference with normalizing flows. In *International conference on machine learning* pages 1530–1538. PMLR.
- [20] Russell , C. T., Luhmann , J. G., & Strangeway , R. J. () Russell , C. T., Luhmann , J. G., & Strangeway , R. J.:
- [21] Timar , A., Opitz , A., Nemeth , Z., Bebesi , Z., Biro , N., Facskó , G., Koban , G., & Madar , A. (2024) 3d pressure-corrected ballistic extrapolation of solar wind speed in the inner heliosphere. *Journal of Space Weather and Space Climate* **14**:14.

- [22] Whittlesey , P. L., Larson , D. E., Kasper , J. C., Halekas , J., Abatcha , M., Abiad , R., Berthomier , M., Case , A. W., Chen , J., Curtis , D. W., Dalton , G., Klein , K. G., Korreck , K. E., Livi , R., Ludlam , M., Marckwardt , M., Rahmati , A., Robinson , M., Slagle , A., Stevens , M. L., Tiu , C., & Verniero , J. L. (2020) The Solar Probe ANalyzersElectrons on the Parker Solar Probe. *The Astrophysical Journal Supplement Series* **246**(2):74.
- [23] Wilson III , L. B., Stevens , M. L., Kasper , J. C., Klein , K. G., Maruca , B. A., Bale , S. D., Bowen , T. A., Pulupa , M. P., & Salem , C. S. (2018) The Statistical Properties of Solar Wind Temperature Parameters Near 1 au. *The Astrophysical Journal Supplement Series* **236**(2):41.