Reconstruction of Time Series Observed in Linear Magnetized Plasma PANTA via a Machine Learning Algorithm

Yasuhiro NARIYUKI, Makoto SASAKI[1)], Tohru HADA[2)] and Shigeru INAGAKI[1)]

Faculty of Human Development, University of Toyama, 3190 Gofuku, Toyama City, Toyama 930-8555, Japan
[1)]*Research Institute for Applied Mechanics, Kyushu University, 6-1 Kasugakoen, Kasuga City, Fukuoka 816-8580, Japan*
[2)]*Faculty of Engineering Sciences, Kyushu University, 6-1 Kasugakoen, Kasuga City, Fukuoka 816-8580, Japan*

(Received 24 April 2019 / Accepted 2 August 2019)

Reconstruction of turbulence time series in a statistically stationary state is discussed by using a machine learning algorithm. We use data obtained by Langmuir probes in the Plasma Assembly for Nonlinear Turbulence Analysis (PANTA). It is shown that even if the distance between two probes is not adequate to resolve the turbulence, the nonlinear regression via the machine learning can give reconstruction better than those by the linear regression and the linear interpolation. Wave forms and frequency spectra show that drift waves are well reconstructed by the machine learning.

Keywords: drift wave, linear magnetized plasma, machine learning, turbulence, coarse-grained model

DOI: 10.1585/pfr.14.1301157

In magnetized plasmas, turbulence often drives structures such as zonal flows and streamers [1], which significantly affect the transport. Thus, the detailed measurement of the turbulence is essentially important. On the other hand, the arrangement of probes are often limited due to the constraint of experimental devices. It is noteworthy that along with developments of computational environments, the machine learning algorithms [2] have received keen attention from many research subjects and fields. Recently, several machine learning techniques have been applied to the solar flare forecasting [3, 4]. It may thus be worthwhile also to examine whether the machine learning methods can successfully be applied to the turbulence time series observed in plasmas. In particular, we ask ourselves whether the methods can correctly reconstruct the turbulence time series at a location where the probe is missing, using the time series at different locations.

In this Rapid Communication, the reconstruction of a probe data of the Plasma Assembly for Nonlinear Turbulence Analysis (PANTA) [5] from the other probe data sets with a supervised machine learning algorithm (random forest [6]) is reported. Thirty-two (32) probes are placed at even intervals in the azimuthal direction of the cylinder. The number of probes ($P = 1, 2, \cdots, 32$) correspond to azimuthal angle difference ($\Delta\theta = \frac{\pi}{16}, \frac{\pi}{8}, \cdots, 2\pi$). We here demonstrate that the standard machine learning technique is applicable to the time series data that is in a statistically stationary state. To predict the time series data measured with the target probes (we choose probe number $P = 15$ and 31), six probes at regular intervals (($P \pm N, P \pm 2N, P \pm 3N$), where N is an integer number) are used

Due to the periodicity, statistical characteristics of time series data obtained by each probe are similar. The data used in this study show the formation of the streamer [7], which is an azimuthal bunching of fluctuation. In this study, the drift waves and the nonlinear quasi-mode (mediator), which has much longer time scale than that of the drift wave, coexist.

In this study, we use time series of electron density (ion saturation current) with the sampling time 10^{-6} s. The data during the time interval$[0.290\,\mathrm{s}, 0.300\,\mathrm{s}]$ are used as the training data of the machine learning. For both the training data and the validation data, the length of the time window is 0.01 s, which is sufficiently longer than time periods of the azimuthal bunching of fluctuation (shown in the frequency spectra of Fig. 4 (b) at about 1.5 kHz). In this study, R language (version 3.5.1) [8] is used to carried out regressions. For multi-variable (multiple) linear regression, fitting linear model in stats package [8] is used. The random forest algorithm [6] is applied to carry out the nonlinear regression via randomForest package [9], which has also been used in recent studies [3, 10]. In this study, hyperparameters in the randomForest package are *nodesize* = 5 (default value), *mtry* = 2 and *ntree* = 1600, respectively. Remark that the results in this study are not sensitive to tuning of the hyperparameters. Such a property of the random forest is also mentioned in the previous study [3], in which the prediction of the solar flare with the machine learning technique was discussed.

Figures 1 and 2 show (a) Pearson's correlation coefficients and (b) sums of squared error (SSE) between original time series of target probes and predicted time

author's e-mail: nariyuki@edu.u-toyama.ac.jp

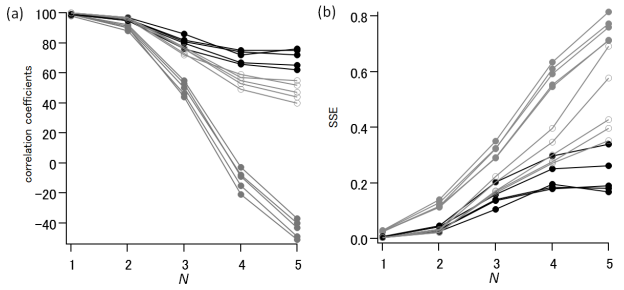

Fig. 1 (a) Pearson's correlation coefficients and (b) sums of squared error (SSE) between original time series of target probes and predicted time series by the random forest [6, 9] (black marks) for $P = 15$. Five data plots at the same N correspond to five time windows ([0.325 s, 0.335 s], [0.350 s, 0.360 s], [0.400 s, 0.410 s], [0.450 s, 0.460 s], [0.500 s, 0.510 s]), respectively. Probes used in the random forest are $(P \pm N, P \pm 2N, P \pm 3N)$, where N is an integer number. For comparison, reconstructions with the multi-variable linear regression (gray non-filled marks) with 6 probes and the linear interpolation (gray filled marks) between two neighboring probes $(P \pm N)$ are also shown.

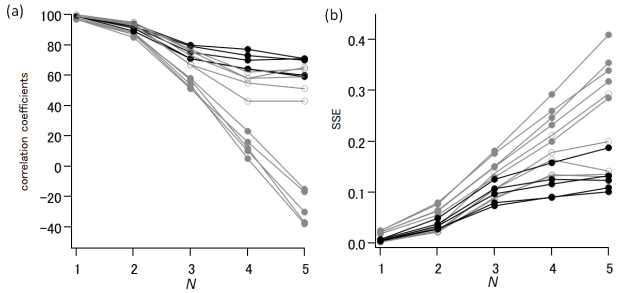

Fig. 2 Same as Fig. 1 except for $P = 31$.

series for $P = 15$ and 31. For comparison, reconstructions with the multi-variable linear regression (gray non-filled marks) with 6 probes and the linear interpolation (gray filled marks) between two neighboring probes $(P \pm N)$ are shown. As shown in Figs. 1 and 2 (a), reconstructions by the linear interpolation with $N = 1, 2$ are fairly accurate, while correlation coefficients decrease with increasing $N(\geq 3)$. It is because the distance between two probes (coarse-graining scale $2N + 1$) is longer than the wavelength of the drift waves [7] when $N \geq 3$. On the other hand, reconstructions by the random forest (black marks) with six probes are much better than those by the simple linear interpolation. This is due to both number of probes and algorithm. The multi-variable linear regression with 6 probes also show better reconstruction than the linear interpolation, while it is worse than those by the random forest when $N = 4$ and 5. This simply indicates that nonlinear regression can give better fitting than the linear regression. On the other hand, when we apply the constructed regression model to data series obtained from other experiments, the random forest algorithm becomes worse than the linear

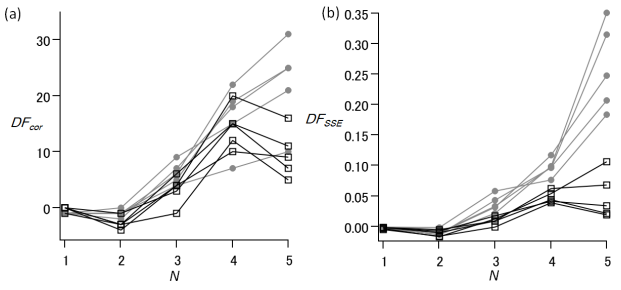

Fig. 3 (a) DF_{cor} and (b) DF_{SSE} vs. N. Circles and squares indicate probe number $P = 15$ and 31, respectively. Five data plots at the same N correspond to five time windows same as those in Fig. 1.

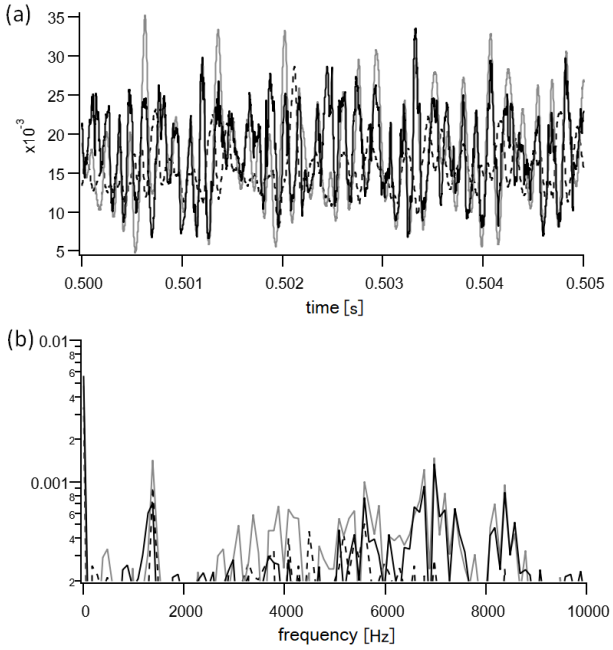

Fig. 4 (a) Time series of probe number $P = 15$ during a time window [0.500 s, 0.510 s] with $N = 4$, where gray-solid, black-solid, and black-dashed lines indicate original data, data predicted by the random forest, and data predicted by the linear interpolation, respectively. (b) Frequency spectra of time series shown in Fig. 4 (a).

interpolation. We revisit this point later.

In order to show the difference between the random forest and the linear regression more clearly, we show

$$DF_{cor} = C_{rf} - C_{lr},$$
$$DF_{SSE} = SSE_{lf} - SSE_{rf},$$

in Fig. 3, where C_{rf} (SSE_{rf}) and C_{lr} (SSE_{lr}) are correlations coefficients (SSE) obtained by using the random forest and the multi-variable linear regression, respectively. In Fig. 3, we find that $DF_{cor} > 0$ and $DF_{SSE} > 0$ in all the present data with $N = 4$ and 5. This suggests that a nonlinear relation in the turbulence is extracted by nonlinear regression via a machine learning algorithm.

Figure 4 (a) shows time series data measured with

probe number $P = 15$ during a time window [0.500 s, 0.510 s] with $N = 4$. The gray-solid, black-solid, and black-dashed lines indicate original data, data predicted by the random forest, and data predicted by the linear interpolation, respectively. As shown in Fig. 4 (a), reconstruction of phases by the random forest is much better than that by the linear interpolation. This simply indicates that data from neighboring probes ($P \pm N$) are not sufficient for the reconstruction but time series data of the other probes ($P \pm 2N$ and $P \pm 3N$)) is necessary when $N \geq 3$. Frequency spectra of time series (Fig. 4 (b)) indicate that the spectra of drift waves (around 5 – 9 kHz) are well reconstructed by the random forest, while those of the mediator (around 1.5 kHz) becomes worse than the linear interpolation. Namely, the high accuracy of the random forest comes from the accurate reconstruction of the drift waves.

Finally we discuss the prediction of the data obtained by the other discharge. Figures 5 (a) and (b) show spatiotemporal evolution of original (training) data, which is discussed above, and test data of the other discharge. Since propagating direction in Fig. 5 (b) is opposite to that of Fig. 5 (a) and the number of probes is 64, probe numbers in Fig. 5 (b) are labeled at intervals of one probe in the order opposite to Fig. 5 (a). To predict data in Fig. 5 (b), both training and test data are normalized (standardized).

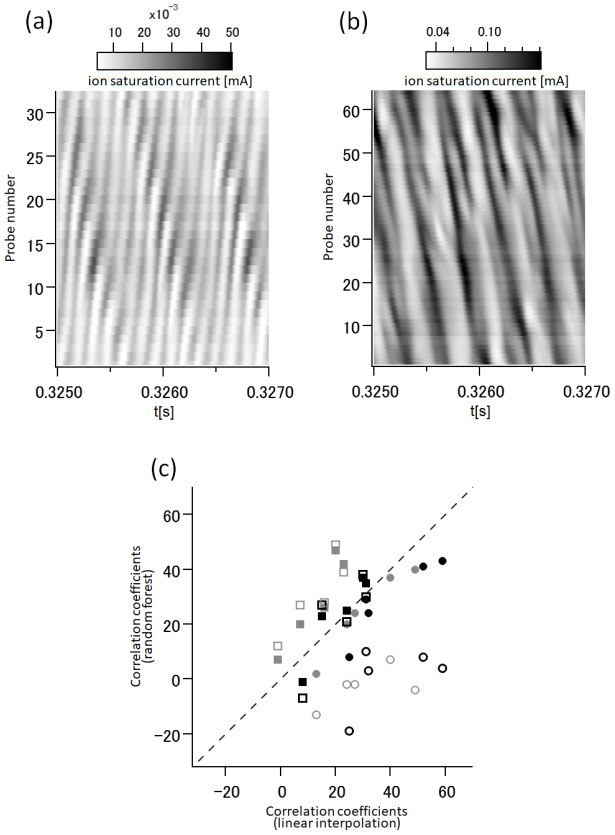

Fig. 5 Spatiotemporal evolution of (a) original data discussed in Figs. 1 and 2 and (b) test data of the other discharge. (c) Pearson's correlation coefficients between time series of test data and predicted time series. The vertical and horizontal axes indicate correlation coefficients with the random forest and the linear interpolation, respectively. Circles and squares indicate probe number $P = 15$ and 31. Black and gray marks correspond to $N = 4$ and 5.

In contrast to data in Fig. 5 (a), the linear interpolation in Fig. 5 (b) is not small even when $N \geq 4$ (Fig. 5 (c)), while predictions of the random forest with a probe ($P = 15$) become much worse than those of the linear interpolation (non-filled circles in Fig. 5 (c)). Filled circles and squares in Fig. 5 (c) show cases which explained data ($P = 15$ and 31) are changed to discrepancies between original training data and predictions of the linear interpolation. These plots clearly show that correlation coefficients shown by circles ($P = 15$) are improved by the change of the definition of the explained data, while some squares ($P = 31$) slightly get worse. Such an unstable behavior of predictions by machine learning is usually due to extrapolation. For instance, correlation coefficients of the linear interpolation indicated by circles ($P = 15$) are higher than those of squares ($P = 31$). In order to improve the accuracy and the stability of the analysis, further calculation with more data obtained by the other discharges is necessary in future.

Recently, recurrent neural networks [11–13], in which past information of the time series is also used for prediction, have been applied to chaotic dynamical systems [11], geomagnetic activity [12] and nuclear fusion [13]. The machine learning algorithm has also been applied to improve turbulence modeling [14, 15]. In this sense, the limitation of number of probes in this study corresponds to the limitation of the spatial resolution in physical modeling. The comprehensive study including topics mentioned above is also necessary in the future.

In summary, the random forest algorithm is applied to time series of turbulence observed in PANTA. It is demonstrated that time series, which are not sufficiently resolved by the coarse-graining scale, can be reconstructed by the random forest with good accuracy. By comparing reconstructed wave forms and frequency spectra with those of original data, it is found that the good accuracy in the random forest regression stems from the reasonably accurate reproduction of the drift waves.

This work was supported in part by the Collaborative Research Program of Research Institute for Applied Mechanics, Kyushu University.

[1] P.H. Diamond, S.-I. Itoh, K. Itoh and T.S. Hahm, Plasma Phys. Control. Fusion **47**, R35 (2005).
[2] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second edition), Springer Series in Statistics, (Springer-Verlag, New-York, 2009).
[3] K. Florios, I. Kontogiannis, S.-H. Park, J.A. Guerra, F. Benvenuto, D.S. Bloomfield and M.K. Georgoulis, Solar Phys. **293**, 28 (2018).
[4] N. Nishizuka, K. Sugiura, Y. Kubo, M. Den, S. Watari and M. Ishii, Astrophys. J. **835**, 156 (2017).

[5] S. Inagaki, T. Kobayashi, Y. Kosuga, S.-I. Itoh, T. Mitsuzono, Y. Nagashima, H. Arakawa, T. Yamada, Y. Miwa, N. Kasuya, M. Sasaki, M. Lesur, A. Fujisawa and K. Itoh, Scientific Rep. **6**, 22189 (2016).
[6] L. Breiman, Machine Learning **45**, 5 (2001).
[7] T. Kobayashi, S. Inagaki, M. Sasaki, Y. Kosuga, H. Arakawa, F. Kin, T. Yamada, Y. Nagashima, N. Kasuya, A. Fujisawa, S.-I. Itoh and K. Itoh, Plasma Fusion Res. **12**, 1401019 (2017).
[8] R. Core Team, see http://www.R-project.org/ for R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2018).
[9] A. Liaw and M. Wiener, R News **2**(3), 18 (2002).
[10] C. Liu, N. Deng, J.T.L. Wang and H. Wang, Astrophys. J. **843**, 104 (2017).
[11] P.R. Vlachas, W. Byeon, Z.Y. Wan, T.P. Sapsis and P. Koumoutsakos, Proc. R. Soc. A **474**, 20170844 (2017).
[12] M.A. Gruet, M. Chandorkar, A. Sicard and E. Camporeale, Space Weather **16**, 1882 (2018).
[13] D.R. Ferreira and JET Contributors, arXiv:1811.00333v1 [physics.plasm-ph] (2018).
[14] J.-X. Wang, J.-L. Wu and H. Xiao, Phys. Rev. Fluids **2**, 034603 (2017).
[15] J.-L. Wu, J.-X. Wang, H. Xiao and J. Ling, Flow Turbulence Combust **99**(1), 25 (2017).