Climate extremes factor attribution: a small data challenge in ML realm

PRASHANT Dave[1]

[1]Indian Institute of Technology Bombay

November 23, 2022

Abstract

The identification of factors driving the climate extremes have been conventionally driven by the physical models evaluated using global climate models and/or using statistical analysis.
However, owing to lack of spatial historical records, both of these approaches pose a data insufficiency challenge. Moreover, identification of primary drivers of climate extremes from a larger set of factors can pose another challenge. Bagging machine learning models in conjugation of synthetic sampling techniques can address both of these challenges.
Here, I demonstrate the applicability of three synthetically sampling techniques along with Random Forest (RF) to identify the main drivers and their spatial locations affecting the heatwave days over India for the period of 1979-2013. The three sampling techniques used to generate balanced data are undersampling, oversampling and synthetic minority oversampling technique (SMOTE). It was RF model with SMOTE that could identify the most important factors with greater precision and recall ($f1-$score (0.85)) as compared to other sampling techniques. Geopotential height\@500 hPa along with sensible heating fluxes were identified as important factors characterizing the Indian heatwave days. The work has repercussion for any of the climate extremes which lacks balanced data along with significantly lesser number of observations than the factors.

Climate extremes factor attribution: a small data challenge in ML realm

Prashant Dave

[1]Center for Climate Studies, Indian Institute of Technology Bombay

Key Points:

- There are lack of historical observations with imbalanced data are available for climate extremes.
- Sampling techniques along with Random Forest can identify the prime drivers of climate extremes.

Corresponding author: Prashant Dave, prashantdave25@gmail.com

Abstract

The identification of factors driving the climate extremes have been conventionally driven by the physical models evaluated using global climate models and/or using statistical analysis. However, owing to lack of spatial historical records, both of these approaches pose a data insufficiency challenge. Moreover, identification of primary drivers of climate extremes from a larger set of factors can pose another challenge. Bagging machine learning models in conjugation of synthetic sampling techniques can address both of these challenges.

Here, I demonstrate the applicability of three synthetically sampling techniques along with Random Forest (RF) to identify the main drivers and their spatial locations affecting the heatwave days over India for the period of 1979-2013. The three sampling techniques used to generate balanced data are undersampling, oversampling and synthetic minority oversampling technique (SMOTE). It was RF model with SMOTE that could identify the most important factors with greater precision and recall ($f1-$score (0.85)) as compared to other sampling techniques. Geopotential height500 hPa along with sensible heating fluxes were identified as important factors characterizing the Indian heatwave days. The work has repercussion for any of the climate extremes which lacks balanced data along with significantly lesser number of observations than the factors.

Plain Language Summary

Understanding the factors characterizing the climate extremes is a challenging task due to lack of observations of climate extremes and interdependence of multiple factors. To address these issues, data can be generated synthetically and bagging methods (a class of machine learning models) can be used to identify the main factors driving the climate extreme. Here, I have demonstrated the applicability of different sampling technique with Random Forest machine learning modeling technique to identify the most important factors characterizing the heatwave over India.

1 Introduction

The identification of factors driving the climate extremes have been conventionally driven by the climate models(Perkins et al., 2012; Mondal et al., 2020; Krishnan et al., 2016; Maharana & Dimri, 2015; Kaufman et al., 2006) and/or statistical analysis (Dave et al., 2020; Rohini et al., 2016; Ratnam et al., 2016; Purnadurga et al., 2018; De et al., 2005; van Oldenborgh et al., 2018; Kodra et al., 2011). Both of these approaches require *a priori* understanding of the underlying physics which subsequently drives the formulation of the hypothesis followed by analysis of the factors to validate the hypotheses. One of issues with these approaches is that there are large number of inter-dependent variables in climate domain and selection of important factors may be subjected to human understanding of the phenomena.

In this regard, purely data driven ML approaches have shown a great potential in enhancing our capability of predicting the extremes events as well as farther our understanding of the underlying mechanisms (Jones, 2017; Ganguly et al., 2014). E.g. O'Gorman and Dwyer (2018) demonstrated potential use of ML to mimic the parameterization of moist convection and modeling climate extremes. Using deep learning researcher (Ham et al., 2019) were able to predict the El-Nino events with over 95% prediction capability. However, application of ML to climate extremes poses its own challenges.

To being with, one of the prime requirement for ML approaches is the large volume of data, that is used to train the model (Jones, 2017). This may be an important issue while we are trying to model extreme events as extremes are not so frequent as compared to nominal days. For example, over India the recorded heatwave events are available for the past 40 years, with sparse heatwave events. The fraction of heatwave days as a fraction to total summer days is very small ($\approx$0.05). This makes data (heatwave and non-heatwave days) imbalanced with large fraction of majority class (where, majority class being non-heatwave days and minority class being heatwave days). This is expected to be another recurrent issue while applying ML techniques to analyze any extreme events. Another challenge which is faced in modeling climate extreme is small data-big data challenge, where data is available spatially but lacks any historical records (Ganguly et al., 2018) and identifying the important factors from a vast set of potential factors becomes challenging. For example, multiple factors are important to characterize heatwave days over India such as geopotential height, latent and sensible heating fluxes, aerosols etc. Moreover, each factor can originate from different location. E.g. latent and sensible affect the heatwave days prediction locally (Rohini et al., 2016) while geopotential height all the way over Africa can play a role in prediction of heatwave days (Ratnam et al., 2016), and the aerosol effect can be locally as well as non-locally (Dave et al., 2020; Mondal et al., 2020). In order to account for factors influence from all the spatial locations, each factor at each location can be considered as a different factor. This increases the total number of factors as compared to the limited available observations.

One of the ways to address these issues of lack of observations, imbalance data and less observations than factors is to increase the minority class data using different sampling techniques such as oversampling, synthetic minority over sampling (SMOTE) (Nitesh V. et al., 2006) etc. Further, using ML techniques such as Random Forest, XGBoost which selects randomly a sub-set of factors and observations for training the model issue of less observations than factors can be addressed. It has been shown that RF model performance does not deteriorate even if the ratio of observation/variables is less than $1/500^{th}$, owing to random sub-sampling of features and observations for each tree in RF.

Here, using the heatwave days data for the period of 1979-2013, I demonstrated the use of sampling techniques Undersampling, Oversampling, SMOTE etc. to address the issue of imbalanced data along with RF modeling approach to take into consideration small observations to factor ratio. Heatwave events are classified as climate extremes and have been witnessed globally(Ding et al., 2010; Perkins et al., 2012) having severe socio-economic impacts on daily lives of people. Over India, a prominent increase in frequency and intensity of heatwave days has been observed (Pai et al., 2013; Rohini et al., 2016; van Oldenborgh et al., 2018). Recently, in the year 2015 one of the rarest and deadliest heatwave events was witnessed across India which resulted in about 2500 deaths (Burton, 2015; Ghatak et al., 2017). This emphasize the importance to enhancing our understanding of the factors affecting heatwave days.

By the analysis, we found that the RF algorithm with SMOTE sampling technique showed best $f1$-score of 0.81 as compared to OVER (0.76) and UNDER (0.49). The RF model could discern the regions of geopotential height 500hPa (GP500) along with regions of latent heat fluxes, sensible heat fluxes, longwave heating and shortwave heating which has been identified to characterize the heatwave over India. Apart from this, the current work also identified that total aerosol along with their origin that are also important factor characterizing the heatwaves.

The flow of paper is as follows: in the next section I describe the data and methodology used for developing the ML model. In the subsequent section we discuss the results and in the last sections I conclude with summarizing the results and repercussion of the study to model climate extremes.

2 Data and Methodology

In the subsequent subsection, the data used for the analysis and methodology used to develop the models has been discussed.

2.1 Data

For the analysis, The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) (Gelaro et al., 2017) data were used for the following variables for the period of March-May (MAM) of 1979-2013 at 5×5 resolution: 1) Geopotential height500hPa (GP500), 2) Greenness Index (GRN), 3) Latent heating land (LHLAND), 4) Sensible heating land (SHLAND), 5) Longwave land (LWLAND), 6) Shortwave land (SWLAND), 7) Black carbon columnar mass (BCCMASS), 8) Black carbon surface mass (BCSMASS), 9) Dust columnar mass (DUCMASS), 10) Dust surface mass (DUSMASS), 11) SO2 columnar mass (SO2CMASS), 12) SO2 surface mass (SO2SMASS), 13) SO4 columnar mass (SO4CMASS), 14) SO4 surface mass (SO4SMASS), 15) Total extinction tau (TOTEXTTAU), 16) Total scattering tau (TOTSCATTAU), and 17) Total angstrom tau (TOTANGRTAU).

The longitude and latitude varied from 0^o to 360^o and -90^o to 90^o, respectively at a resolution of $5^o \times 5^o$. Thus, For each variable for a given longitude and latitude a different variable is considered for the analysis. So there were total 45288 ($17\times37\times72$) variables were used for the analysis.

The heatwave days used for the analysis were obtained as listed in (Dave et al., 2020). If a particular day corresponded to heatwave event it was marked to class '1' and if the day belonged to nominal days it was marked to class '0'. There were total 239 heatwave days out of total 4148 days in the 34 years time period of 1979-2013. The total fraction of heatwave days were ≈ 0.05 (239/4148).

2.2 System configuration

The analysis was performed on an Intel(R) Core(TM) i5-8250U CPU with 1.60GHz, 4 Cores and 8 Logical Processors system.

2.3 Methodology

Given the small number of heatwave days as compared to total days we either need to decrease the majority class observations (undersampling) or increase the observation of minority class (oversampling and SMOTE).Further, I used RF to develop ML model. The choice of RF methodology was motivated due to less number of observations as compared to the total number of variables i.e ≈ 0.09 (4148/45288).

2.3.1 Sampling techniques and evaluation of the model

For the current analysis, we have compared the performance of the model using undersampling, oversampling and SMOTE sampling techniques.

Undersampling (UDNER): In this process random observations from the majority class are removed to match the observations in minority class.

Oversampling (OVER): In this process random observations from the minority class are added to match the observations in majority class.

Synthetic minority oversampling technique (SMOTE): In SMOTE (Nitesh V. et al., 2006) sampling technique, synthetic observations from the minority class are generated using k-nearest neighbors to match the observations in the majority class.

2.3.2 Modeling methodology, evaluation metrics and factor score

Once the imbalanced-data was transformed into balanced-data, data were split into training data and testing data with 80:20 ratio. Using training data, RF technique was used to develop the model with the objective of predicting the heatwave days with high $f1$-score and precision-recall curve. $f1-$score is the harmonic mean of precision and recall. In case of imbalanced data, $f1$-score and precision-recall curve are better predictor of model performance as compared to accuracy and AUC-ROC.

The RF modeling techniques randomly sub-sample the observations and feature for each tree in the forest. Thus, the RF is a suitable technique when the variables are more as compared to observations. The RF model was fine tuned using Cross-validation (CV) for each of the model developed using data generated with different sampling techniques. The parameters that were tuned are n_estimators: number of trees, max_depth: maximum depth of tree i.e. maximum depth between root node and minimum_samples _split: minimum number of samples needed at a node for split (Table 1). The following other hyper-parameters were kept same for all the three sampling techniques: random_state=0; min_samples_leaf=1; n_jobs=3; min_weight_fraction_leaf=0; min_impurity_decrease=0; max_feature='auto'.

Table 1. Hyper-parameters for different sampling techniques

Sampling technique	n_estimators	max_depth	min_samples_split
UNDER	2500	20	2
OVER	2400	7	10
SMOTE	900	20	2

Once the model was I identified the most important factors which are playing significant role in increasing the predictive power of the model. These factors are identified using the factor score, which is a relative score assigned to all the factors used for the modeling. For each factor one score was assigned by each of the three models. In order to compare the scores across different sampling techniques based RF models scores were scaled between 0 and 1 using following transformation equation:

$$Score_i = \frac{Score_i - Score_{min}}{Score_{max} - Score_{min}} \tag{1}$$

Here, $Score_i$ is the score of the ith factor, $Score_{max}$ is the maximum score across of the factors and $Score_{min}$ is the minimum score across of the factors.

3 Results and Discussion

3.1 Sampling methods

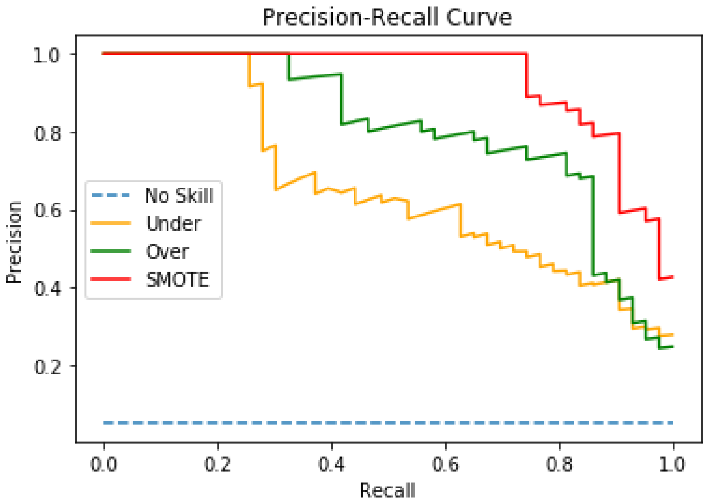

Figure 1. Comparison of UNDER, OVER and SMOTE sampling techniques

In the Figure 1, precision-recall curves are depicted for the models with three sampling techniques. In the precision-recall curve, greater the area-under-the-curve larger is the discriminatory power of the model. Here, class '0' represents no heatwave day and class '1' represents a heatwave day. "No Skill" is where all the observations in test data are classified to either of the class using random guess, therefore each class has the probability of 0.5. In Table 2, the threshold used to differentiate between class 0 and 1 is listed for each of the sampling techniques. These thresholds are identified from the precision-recall curve (Figure 1), where we have largest precision and recall. If the probability is below the threshold, class is assigned as 0 (non heatwave day) otherwise 1 (heatwave day). The $f1$-score is listed in Table 2. We see that $f1$-score for UNDER is lower than OVER, which is lower than SMOTE. The SMOTE sampling algorithms shows highest area under the curve and thus, exhibit largest discriminatory powers. This implies that the SMOTE sampling technique can identify the factors that can differentiate between heatwave and non-heatwave days to a greater extent as compared to other sampling techniques.

Table 2. Classification report for UNDER, OVER and SMOTE sampling technique

Technique	Threshold	Class	precision	recall	$f1$-score	support
UNDER	0.389	0	0.98	0.98	0.98	787
		1	0.60	0.60	0.60	43
OVER	0.657	0	0.99	0.99	0.99	787
		1	0.74	0.74	0.74	43
SMOTE	0.476	0	0.99	0.99	0.99	787
		1	0.86	0.84	0.85	43

In UNDER sampling there is information loss as the observations from majority class are dropped while the OVER sampling does not add any new information as the observations from minority class are repeated randomly. However, in SMOTE new samples are generated using the nearest neighbor approach that adds variability to the existing observations (Nitesh V. et al., 2006). This could be one the reasons for high $f1-$score obtained with SMOTE. Further, the differentiation between heatwave and non-heatwave days depends upon the score assigned to different variable and subsequently I present and discuss data and score assigned by SMOTE, OVER and UNDER sampling to different variables.

In Figures 2(a-c), geopotential height 500hPa anomaly (GP500) averaged across heatwave days identified by SMOTE, OVER and UNDER sampling are shown. It can be seen that all the three sampling techniques identify a high GP500 anomaly over Indian subcontinent. In Figures 2(d-f) the score assigned to each spatial location by SMOTE, OVER and UNDER sampling methods are shown. From the Figures 2(d-f), it can be noted that SMOTE sampling technique (Figure 2d)) assigns large weight (>0.75) to GP500 anomaly as compared to OVER and UNDER sampling techniques over Indian region. Further, there is an extension of GP500 anomaly over the African region, which has been assigned larger weights in OVER (Figure 2(f)) and UNDER (Figure 2(e)) sampling technique as compared to SMOTE ((Figure 2(d))). Studies (Ratnam et al., 2016; Rohini et al., 2016) have also identified large positive anomaly of GP500 during heatwave event over India. This is owing to development of high pressure conditions with increased atmospheric stability (Ratnam et al., 2016; Rohini et al., 2016). Further, the positive anomaly of GP500 was reported to be extended all the way upto Africa, owing to the Rossby wave source anomalies(Ratnam et al., 2016).

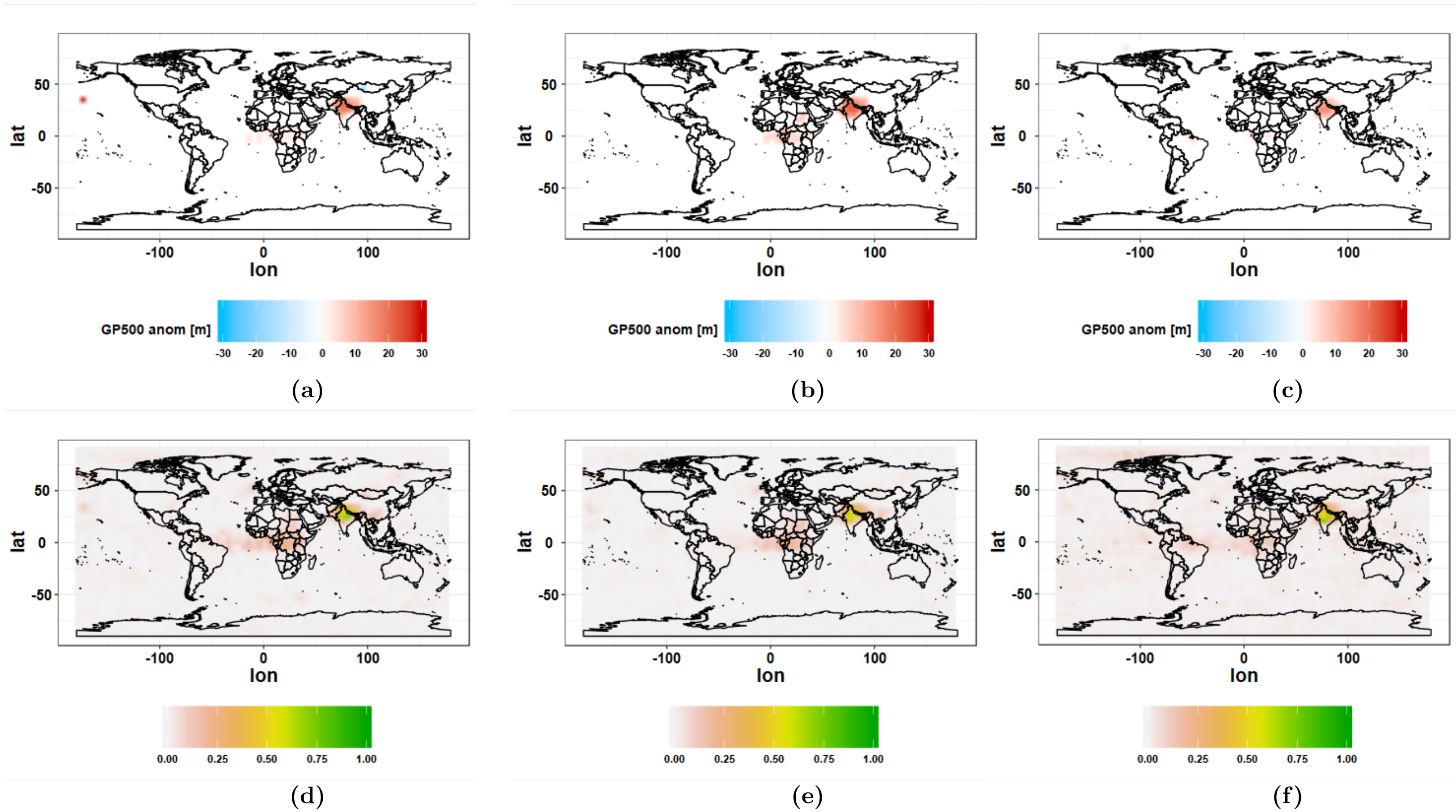

Figure 2. GP500 anomaly magnitude associated with heatwave days and score for (a) SMOTE, (b) OVER and (c) UNDER sampling techniques. (It is to be noted that for simplicity in (a-c) regions corresponding to top hundred score value are presented. In (d-f) all the scores are presented.)

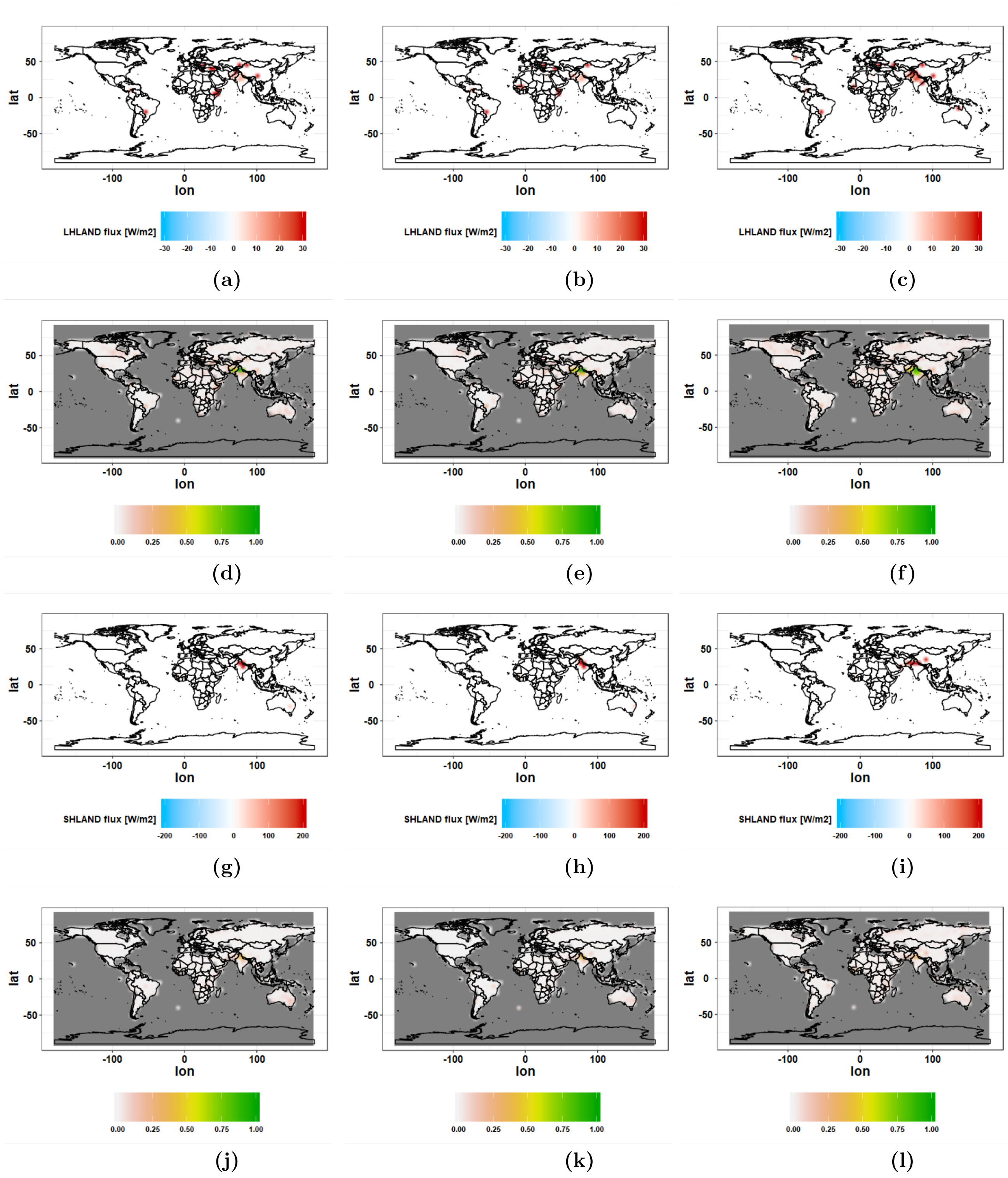

Figure 3. Latent and sensible heating fluxes magnitude and score associated with heatwave days for (a) SMOTE, (b) OVER and (c) UNDER sampling techniques. (It is to be noted that for simplicity in (a-c) and (g-i) regions corresponding to top hundred score value are presented. In (d-f) and (j-l) all the scores are presented.)

Other characteristic features of the heatwave events over India are reduced latent heating fluxes along with increased sensible heating fluxes (Mondal et al., 2020), which can be attributed to low moisture content over the Indian subcontinent during the summer months(Mondal et al., 2020). From the Figure 3(a-c), it can be seen that the latent heating fluxes are small during the heatwave events predicted by SMOTE and OVER sampling techniques while it is larger for the UNDER sampling technique. Also, from the Figure 3(d-f), it can be seen that the region which has been identified as playing a role in affecting the heatwave events over India is larger for UNDER sampling as compared to SMOTE and OVER sampling. This implies that although all the three sampling techniques can identify the region and latent heat flux magnitude, the SMOTE and OVER sampling capture the spatio-temporal variability in a better way as compared to UNDER sampling. It can also be noted that in Figure 3(a-c), there are some regions over North, East and West Africa and Central-South America, showing large magnitude of latent heating flux, however these regions are not assigned significant score (Figure 3(d-f)) and could be due to numerical artifacts.

Similar arguments can be presented for sensible heating fluxes magnitude (Figure 3(g-i)) and score (Figure 3(j-l)). However, it is to be noted that during heatwave events increased sensible heating fluxes have been observed over India (Mondal et al., 2020), which are also reflected in Figure 3(g-i). Further, it can be seen that in UNDER a larger region has been identified (Figure 3(l)), from where sensible heating flux can affect the prediction of heatwave events, as compared to SMOTE (Figure 3(j)) and OVER (Figure 3(k)) sampling techniques. Further larger magnitude of sensible heating is predicted by SMOTE (Figure 3(g)) and OVER (Figure 3(h)) as compared to UNDER (Figure 3(i)). Further, we analyzed the magnitude and score of longwave and shortwave anomaly spatio-temporal distribution predicted by SMOTE, OVER and UNDER sampling techniques.

Heatwave days over India are associated with increased outgoing longwave radiation spread over the North-Western India (Rohini et al., 2016). Here, SMOTE technique identified a large negative (outward direction negative) anomaly in long wave radiation over the North-Western India (Figure 4(a)). Along with this although, OVER and UNDER sampling techniques also show a large negative anomaly of longwave radiation fluxes, the region is spread all the way to southern India (Figure 4(b-c)). Further, the score assigned to longwave fluxes is higher over North-West India while a lower score is assigned as we move away from North-West region by all the three sampling techniques (Figure 4(d-f)).

From the Figure 4(g-i), it can be seen that the positive shortwave heating anomaly is found to be associated with predicted heatwave days over the North India. Here, again the region is found to be spread-out in UNDER (Figure 4(i)). It is to be noted that the score assigned to shortwave heating (Figure 4(j-l)) is low as compared to longwave heating (Figure 4(d-f)). It implies that long wave heating is a better feature that can distinguish between heatwave and non-heatwave days. The shortwave anomaly is high during the whole summer while during the heatwave days persistent clear sky conditions are observed (Rohini et al., 2016) leading to large outgoing radiative fluxes (Rohini et al., 2016) and this could be the reason behind less importance is given to shortwave heating fluxes as compared to longwave fluxes.

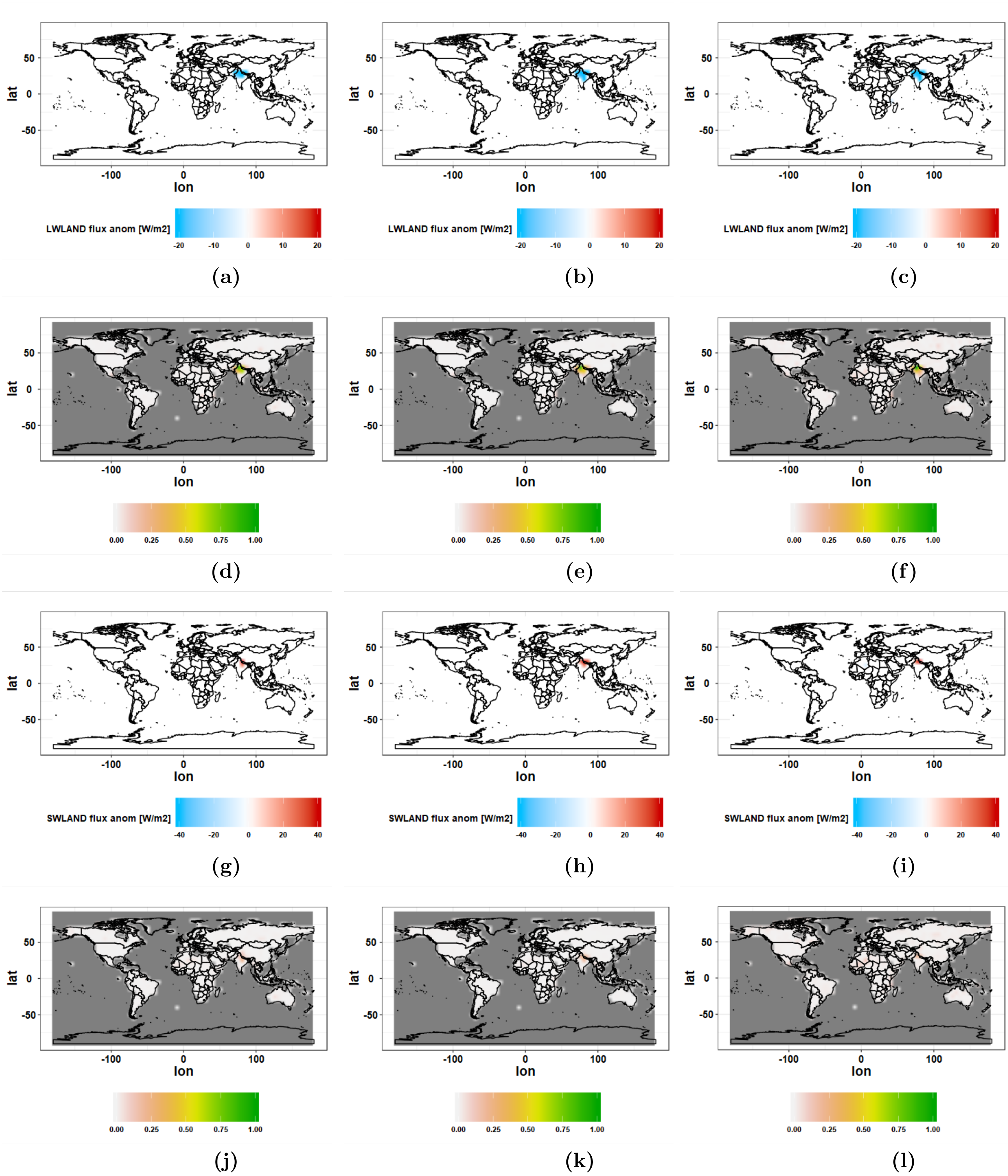

Figure 4. Longwave and shortwave anomaly magnitude and score associated with heatwave days for (a) SMOTE, (b) OVER and (c) UNDER sampling techniques. (It is to be noted that for simplicity in (a-c) and (g-i) regions corresponding to top hundred score value are presented. In (d-f) and (j-l) all the scores are presented.)

Further, the SMOTE predicts high TOTEXT and TOTSCAT over north India and Western Africa (Figure 5(a,g)). While, OVER sampling (Figure 5(b,h)) associates Western Africa and UNDER sampling (Figure 5(c,i)) associates North India, South India and Western Africa with TOTEXT and TOTSCAT. This indicates that SMOTE can identify regions pertinent to heatwave days. The role of local as well as non-local aerosols in exacerbating heatwave conditions over India have been identified by different studies (Mondal et al., 2020; Dave et al., 2020). Although the extinction and scattering due to aerosols have not been assigned a large score as compared to GP500 anomaly, latent and sensible heating fluxes, and longwave and shortwave fluxes, the emergence of region all the way to West Africa does require further investigation. The reason behind this observation could be associated with the presence of anomalous anti-cyclone conditions as a part of a quasi-stationary wave extending all the way upto North-western Africa (Ratnam et al., 2016), which increases the dust aerosols anomaly.

From the Figure 6, we can see that a larger dust anomaly is also identified over the West Africa by SMOTE (Figure 6(a)) and OVER (Figure 6(b)) sampling techniques. This indicates the accumulation of large dust anomalies over Western Africa which can be the result of large TOTEXT and TOTSCAT observed in Figure 5(a-c). This can subsequently can be associated with the observed heatwave days over India and can be a discerning factor. However, UNDER sampling technique (Figure 6(c)) does not capture any dust anomaly over the West Africa.

There are some regions identified near the Northern Australia exhibiting score in the range of 0.1-0.25 (Figure 6(d-f)), which could be an artifact owing to i) small score as compared to factors discussed earlier and ii) the presence of dust is very low in this region (Figure 6(a-c)). However, this is one the factor that requires further investigation, although not focus of this paper.

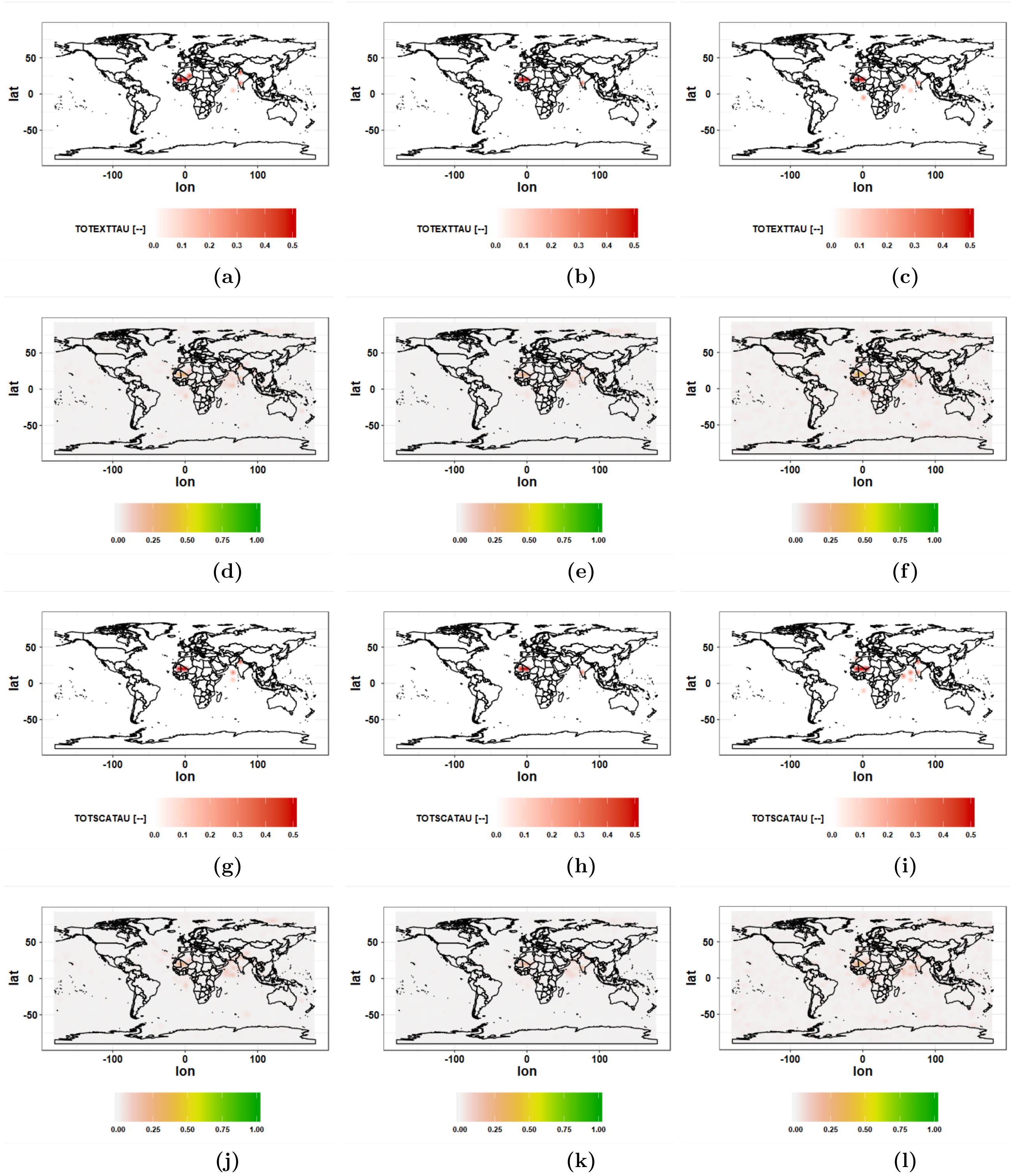

Figure 5. Total extinction and total scattering magnitude and score associated with heatwave days for (a) SMOTE, (b) OVER and (c) UNDER sampling techniques. (It is to be noted that for simplicity in (a-c) and (g-i) regions corresponding to top hundred score value are presented. In (d-f) and (j-l) all the scores are presented.)

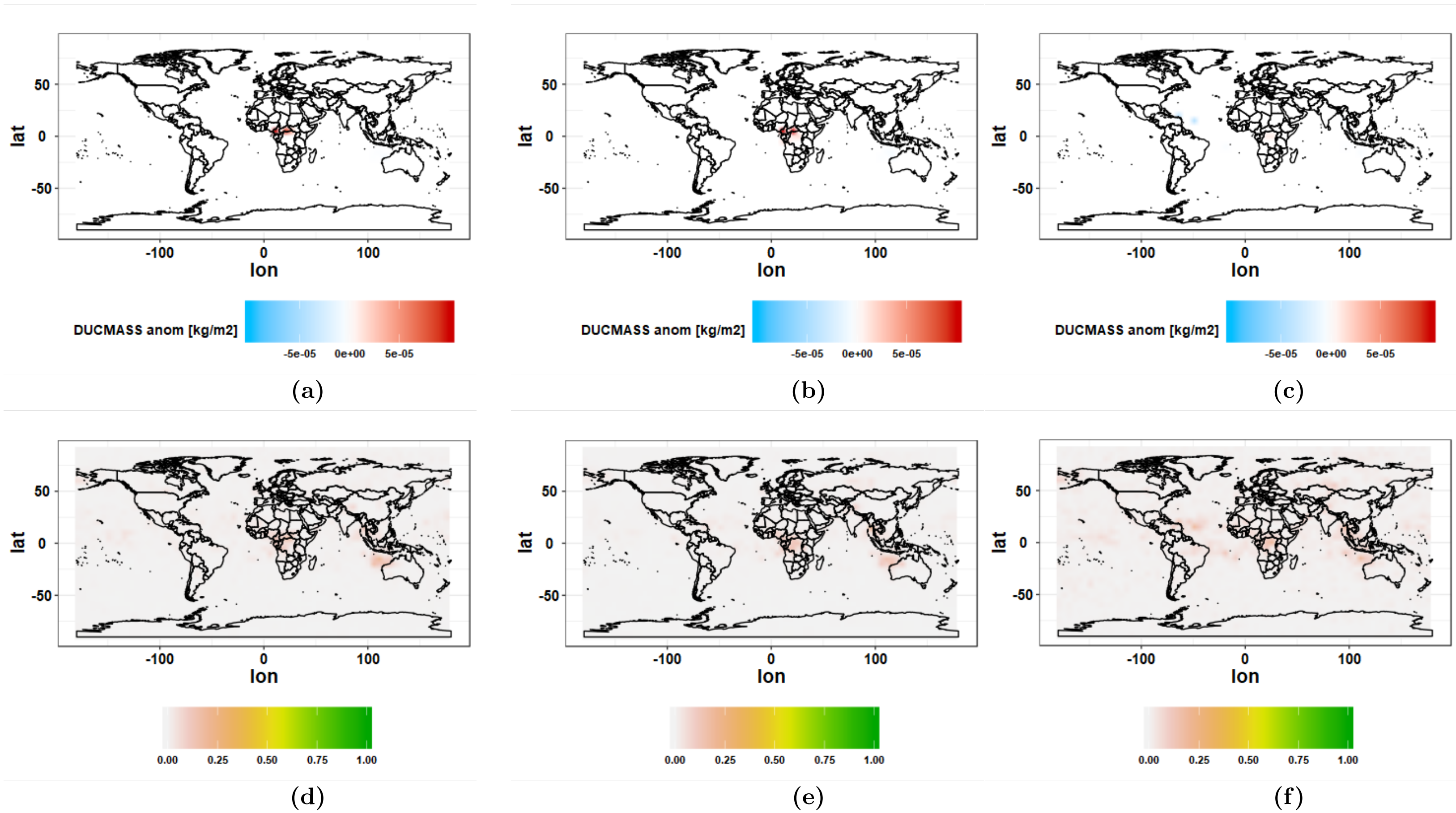

Figure 6. Dust columnar mass anomaly (DUCMASS) magnitude and score associated with heatwave days for (a) SMOTE, (b) OVER and (c) UNDER sampling techniques. (It is to be noted that for simplicity in (a-c) regions corresponding to top hundred score value are presented. In (d-f) all the scores are presented.)

The other factors, i.e. dust surface mass anomaly (DUSMASS, (Figure S1)), Greenness Index (GRN, (Figure S2)), black carbon columnar mass anomaly (BCCMASS, Figure S3 (a-c)), black carbon surface mass anomaly (BCSMASS, Figure S3(d-f)), SO2 columnar mass anomaly (SO2CMASS, Figure S4 (a-c)), SO2 surface mass anomaly (SO2SMASS, Figure S4(d-f)), SO4 columnar anomaly (SO4CMASS, Figure S5(a-c)), SO4 surface mass anomaly (SO4SMASS, Figure S5 (d-f)) and total angstrom (TOTANGSTR, Figure S6) were assigned low score (<0.1) by all the three sampling techniques. The score distribution for these factors are shown in the supplementary information.

This highlights of the limitations of the model is that while it can identify the cumulative effect of aerosol on heatwave days, it could not differentiate between the effect of absorbing and scattering aerosols. This could be due to either small effect of aerosols as compared to other factors.

4 Conclusion and discussions

The analysis of extremes has been constrained by availability of observations. Recent progress in climate modeling has helped significantly in understanding the factors that may play role in characterizing the climate extremes. However, climate models have their own limitations such as parameterization schemes, logistics and resources associated with running a climate model. Here, we presented an alternate approach that uses RF with limited imbalanced observations of heatwave events over India to identify the important factors that can characterize the extreme events. The imbalanced data were transformed into balanced data using SMOTE, OVER and UNDER sampling techniques.

It was found that SMOTE sampling technique performs better (high $f1-$score)) as compared to OVER and UNDER sampling approaches. This can be attributed to generation of new samples in SMOTE using nearest neighbor as compared to repetition of information from minority class (OVER) and of loss of information from majority class (UNDER) sampling. The SMOTE algorithm could identify the important spatial position of factors, e.g. geopotential height, latent and sensible heating, longwave and shortwave fluxes etc., that can delineate between heatwave and non-heatwave days to a larger extent.

In future, the machine learning model performance can be further improved/compared with boosting approaches (such as XGBoost), which have shown better predictive power than RF, as bagging techniques generate trees sequentially using information from previous trees. Overall, the analysis has shown an alternate method to understand the climate extremes with limited data using RF approach with synthetically generated samples. Along with this, such type of modeling does not take much cpu time in identifying drivers of climate extremes along with their spatial distribution and therefore can be easily scaled up.

Acknowledgments

I would like to thank Prof. Mani Bhushan and IIT Bombay for allowing me to use the computing facilities at CAD Center. I would also like to acknowledge R Core Team (2018): A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (http://www.R-project.org/) and python-library scikit-learn, using which the whole analysis was performed. Further, I would like to acknowledge NOAA/OAR/ESRL PSD, Boulder, Colorado, USA for providing the reanalysis data.

References

Burton, C. (2015). *India's deadly heatwave nears end as monsoon arrives.* Retrieved 2018-08-09, from `https://www.theweathernetwork.com/uk/news/articles/indias-deadly-heatwave-nears-end-as-monsoon-arrives/52420/`

Dave, P., Bhushan, M., & Venkataraman, C. (2020). Absorbing aerosol influence on temperature maxima: An observation based study over India. *Atmospheric Environment*, *223*(117237). doi: 10.1016/j.atmosenv.2019.117237

De, R., Dube, K., & Rao, G. S. P. (2005). Extreme weather events over India in the last 100 years. *Journal of Indian Geophysical Union*, *9*(3), 173–187.

Ding, T., Qian, W., & Yanb, Z. (2010). Changes in hot days and heat waves in China during 1961-2007. *International Journal of Climatology*, *30*(10), 1452–1462. doi: 10.1002/joc.1989

Ganguly, A., Kodra, E., Agrawal, A., Banerjee, A., Boriah, S., Chatterjee, S., ... Wuebbles, D. (2014). Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques. *Nonlinear Processes in Geophysics*, *21*(4), 777–795. Retrieved from `https://www.nonlin-processes-geophys.net/21/777/2014/` doi: 10.5194/npg-21-777-2014

Ganguly, A., Kodra, E., Bhatia, U., Warner, E. M., Duffy, K., Banerjee, A., & Ganguly, S. (2018). *Data-driven solutions - Climate2020.* Retrieved 2020-03-09, from `https://www.climate2020.org.uk/data-driven-solutions/`

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., ... Zhao, B. (2017, 06). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419-

5454. Retrieved from https://doi.org/10.1175/JCLI-D-16-0758.1 doi: 10.1175/JCLI-D-16-0758.1

Ghatak, D., Zaitchik, B., Hain, C., & Anderson, M. (2017). The role of local heating in the 2015 Indian Heat Wave. *Scientific Reports*, *7*(1), 1–8. Retrieved from http://dx.doi.org/10.1038/s41598-017-07956-5 doi: 10.1038/s41598-017-07956-5

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, *573*(7775), 568–572. Retrieved from https://doi.org/10.1038/s41586-019-1559-7 doi: 10.1038/s41586-019-1559-7

Jones, N. (2017). Machine learning tapped to improve climate forecasts. *Nature*, *548*(August), 379.

Kaufman, Y. J., Boucher, O., Tanré, D., Chin, M., Remer, L. A., Takemura, T., … Schubert, G. (2006). Sensitivity of precipitation extremes to radiative forcing of greenhouse gases and aerosols. *Geophysical Research Letters*, *32*(7), 1–4. Retrieved from http://doi.wiley.com/10.1002/2016GL070869 doi: 10.1002/2016GL070869

Kodra, E., Chatterjee, S., & Ganguly, A. R. (2011). Exploring Granger causality between global average observed time series of carbon dioxide and temperature. *Theoretical and Applied Climatology*, *104*(3-4), 325–335. doi: 10.1007/s00704-010-0342-3

Krishnan, R., Sabin, T. P., Vellore, R., Mujumdar, M., Sanjay, J., Goswami, B. N., … Terray, P. (2016). Deciphering the desiccation trend of the South Asian monsoon hydroclimate in a warming world. *Climate Dynamics*, *47*(3), 1007–1027. doi: 10.1007/s00382-015-2886-5

Maharana, P., & Dimri, A. P. (2015). Study of intraseasonal variability of Indian summer monsoon using a regional climate model. *Climate Dynamics*, *46*(3), 1043–1064. doi: 10.1007/s00382-015-2631-0

Mondal, A., Sah, N., Sharma, A., Venkataraman, C., & Patil, N. (2020). Absorbing aerosols and high-temperature extremes in india: A general circulation modelling study. *International Journal of Climatology*, *n/a*(n/a), 1-20. Retrieved from https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.6783 doi: 10.1002/joc.6783

Nitesh V., C., Kevin W., B., Lawrence O., H., & W. Philip, K. (2006). snopes.com: Two-Striped Telamonia Spider. *Journal of Artificial Intelligence Research*, *2009*(Sept. 28), 321–357. Retrieved from https://arxiv.org/pdf/1106.1813.pdf{\%}0Ahttp://www.snopes.com/horrors/insects/telamonia.asp doi: 10.1613/jair.953

O'Gorman, P. A., & Dwyer, J. G. (2018). Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events. *Journal of Advances in Modeling Earth Systems*. doi: 10.1029/2018MS001351

Pai, D., Nair, S. A., & Ramanathan, A. (2013). Long term climatology and trends of heat waves over India during the recent 50 years (1961-2010). *Mausam*, *64*(4), 585–604.

Perkins, S. E., Alexander, L. V., & Nairn, J. R. (2012). Increasing frequency, intensity and duration of observed global heatwaves and warm spells. *Geophysical Research Letters*, *39*(20), 1–5. doi: 10.1029/2012GL053361

Purnadurga, G., Lakshmi Kumar, T. V., Koteswara Rao, K., Rajasekhar, M., & Narayanan, M. S. (2018). Investigation of temperature changes over India in association with meteorological parameters in a warming climate. *International Journal of Climatology*, *38*(2), 867–877. doi: 10.1002/joc.5216

Ratnam, J. V., Behera, S. K., Ratna, S. B., Rajeevan, M., & Yamagata, T. (2016). Anatomy of Indian heatwaves. *Scientific reports*, *6*, 1-11. Retrieved from `http://dx.doi.org/10.1038/srep24395` doi: 10.1038/srep24395

Rohini, P., Rajeevan, M., & Srivastava, A. K. (2016). On the variability and increasing trends of Heat waves over India. *Scientific Reports*, *6*, 26153. Retrieved from `http://www.nature.com/articles/srep26153` doi: 10.1038/srep26153

van Oldenborgh, G. J., Philip, S., Kew, S., van Weele, M., Uhe, P., Otto, F., ... AchutaRao, K. (2018). Extreme heat in india and anthropogenic climate change. *Natural Hazards and Earth System Sciences*, *18*(1), 365–381. Retrieved from `https://www.nat-hazards-earth-syst-sci.net/18/365/2018/` doi: 10.5194/nhess-18-365-2018

Supporting Information for "Climate extremes factor attribution: a small data challenge in ML realm"

Prashant Dave

[1]Center for Climate Studies, Indian Institute of Technology Bombay

Contents of this file

1. Figures S1 to S6

Introduction This file contains score assigned by SMOTE, OVER and UNDER sampling techniques to following variables:

1. DUSMASS
2. GRN index
3. BCCMASS
4. BCSMASS
5. SO2CMASS
6. SO2SMASS
7. SO4CMASS
8. SO4SMASS
9. TOTANGSTR

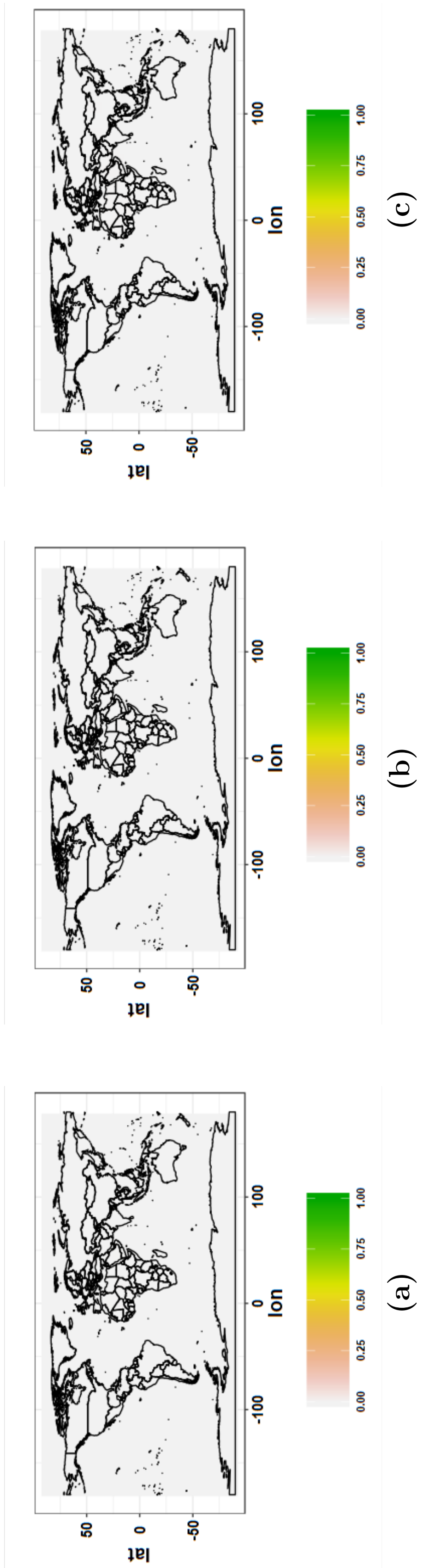

Figure S1. DUSMASS anomaly score for SMOTE, OVER and UNDER sampling techniques

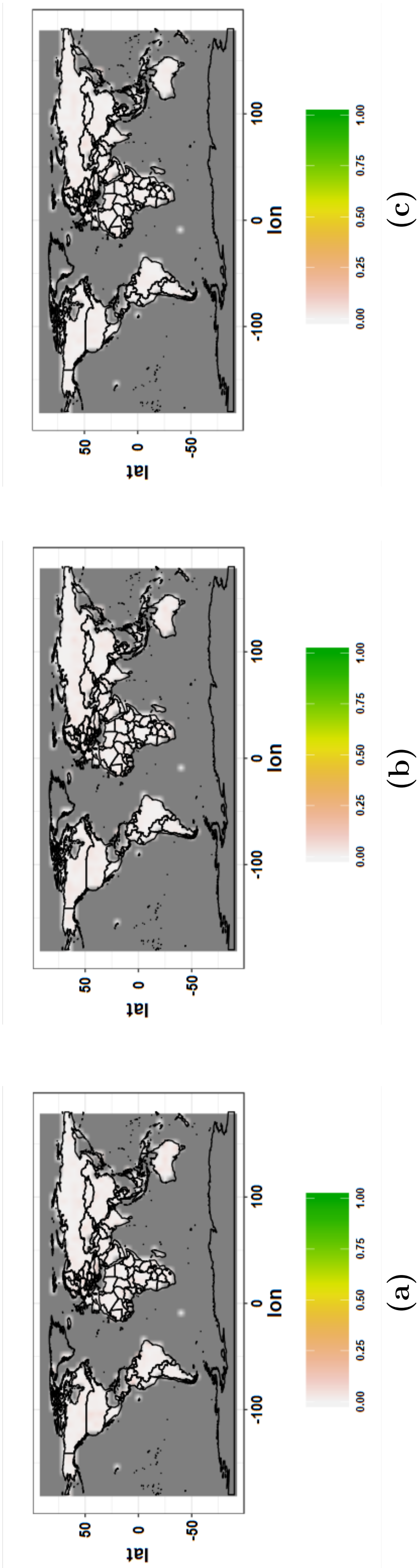

Figure S2. GRN Index score for SMOTE, OVER and UNDER sampling techniques

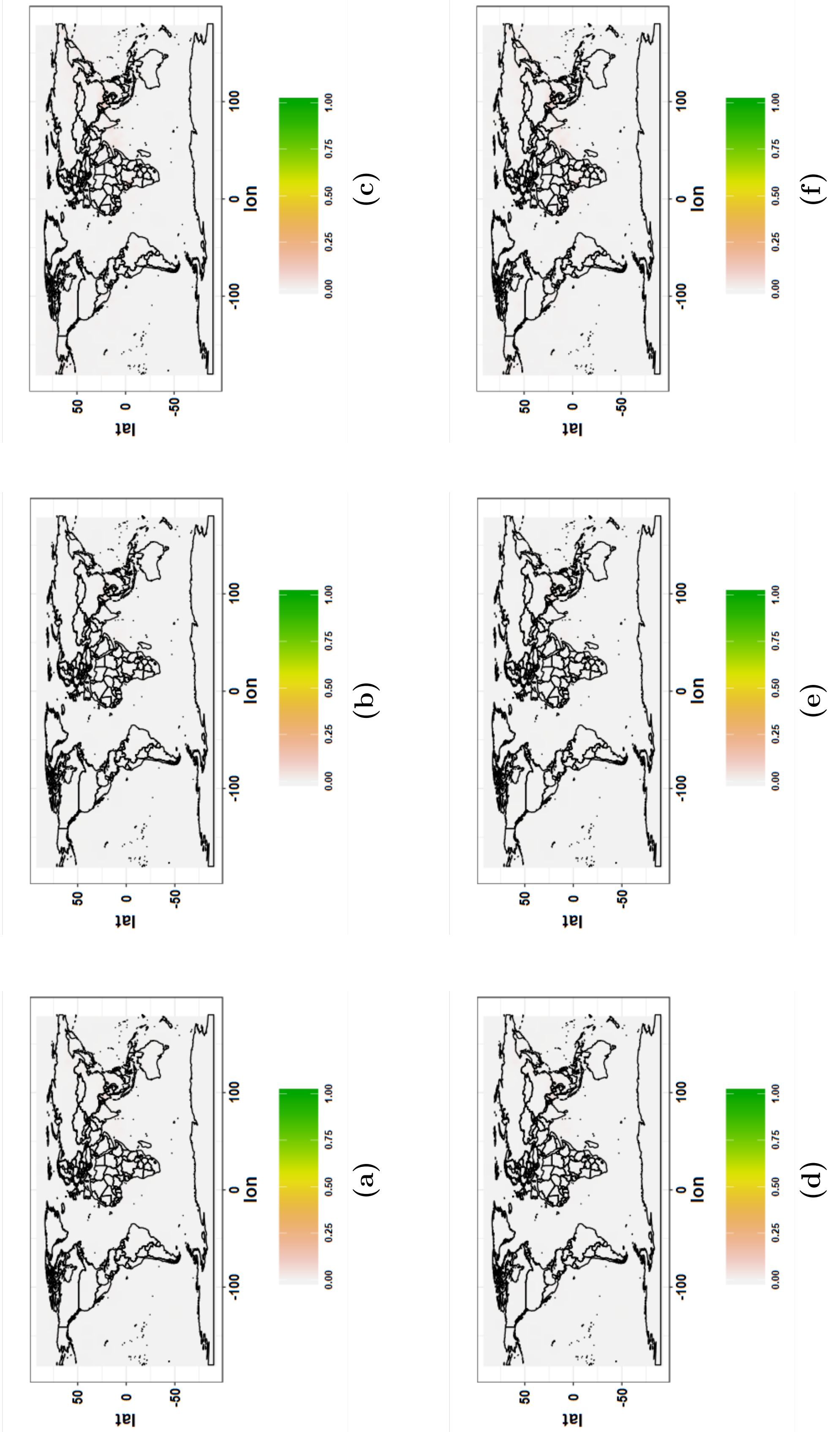

Figure S3. BCCMASS anomaly and BCSMASS anomaly score for SMOTE, OVER and UNDER sampling techniques

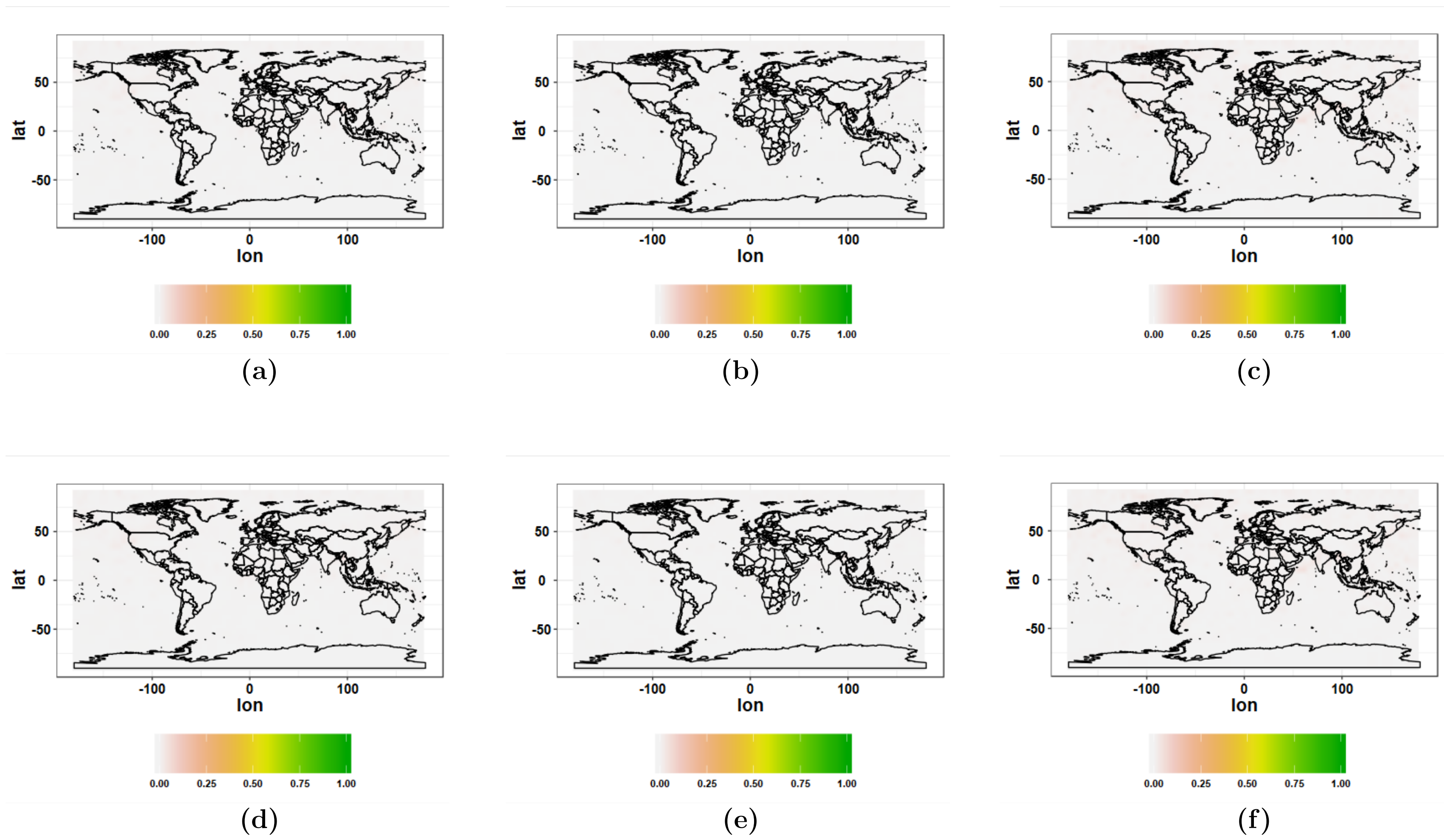

Figure S4. SO2CMASS anomaly and SO2SMASS anomaly score for SMOTE, OVER and UNDER sampling techniques

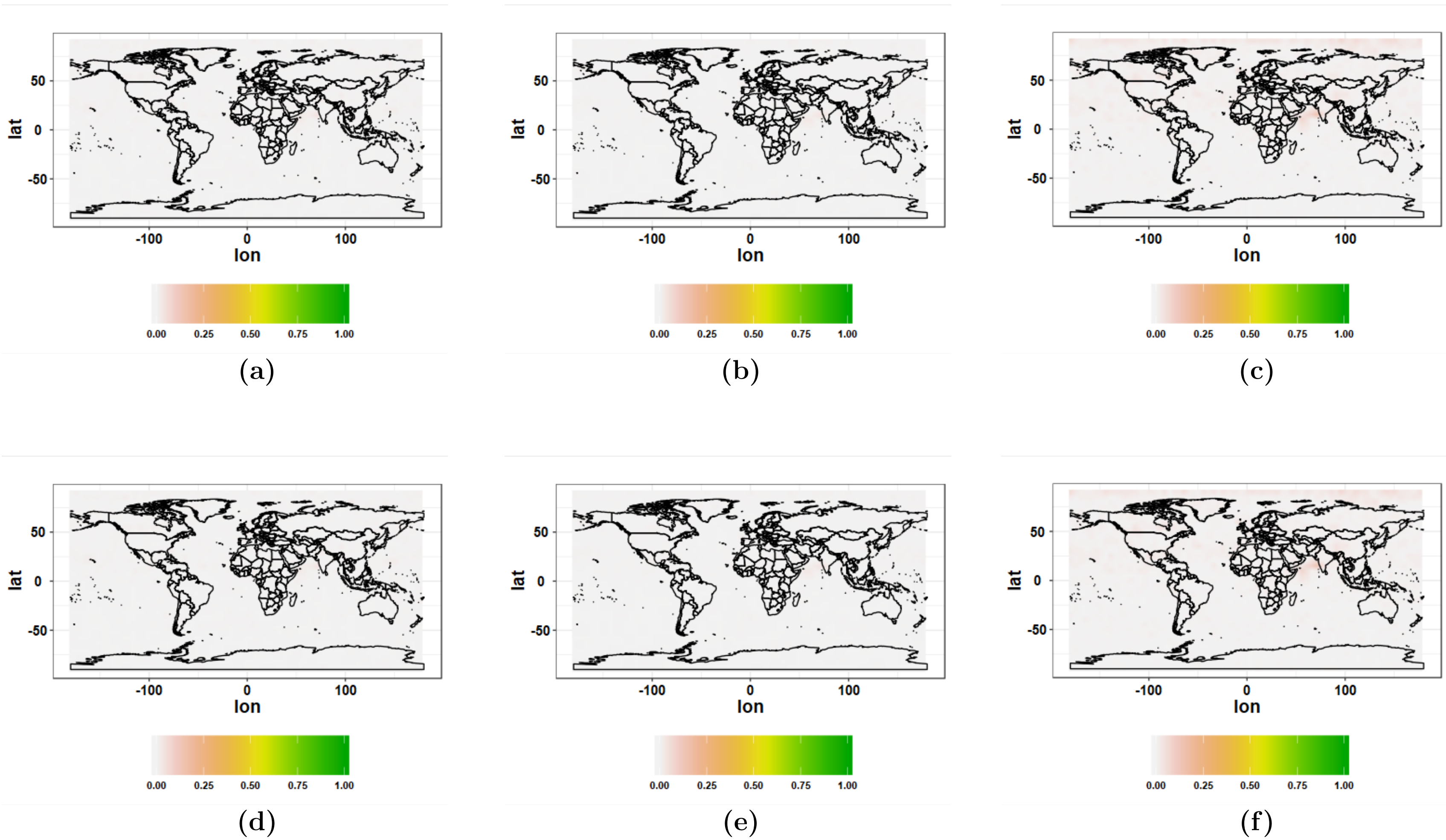

Figure S5. SO4CMASS anomaly and SO4SMASS anomaly score for SMOTE, OVER and UNDER sampling techniques

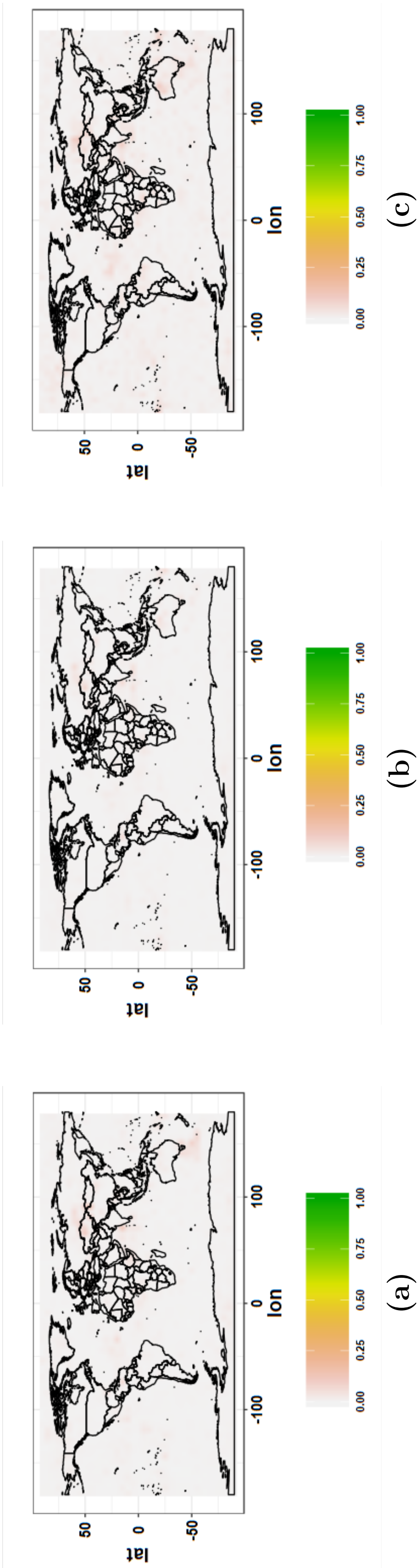

Figure S6. TOTANGSTR score for SMOTE, OVER and UNDER sampling techniques