Assignment Three

Miles Kent

December 18, 2023

Abstract

The purpose of this assignment was to get the first set of full results for the heatwave project. To elaborate, generate ROC curves and calculate model effectiveness through various statistical parameters. I will discuss how we prepared the data for our experiment. I will also discuss our results as well as the necessary back group information needed to calculate our statistics and understand ROC curves and AUC.

Comment

All codes and significant outputs that are not included in this write up are included in the Assignment 3 folder. If you have any questions regarding this document or the code within the Assignment 3 folder, please reach out to me.

1 Data

1.1 Climate Norms

Climate norms are calculated from the daily temperature measurements (see Assignment 1) from the National Oceanic and Atmospheric Administration (NOAA). We calculate norms for each city using data within the window of Jan 1st, 1900 - Dec 31st, 2019. We use the maximum and minimum norms calculated to identify our heatwaves (see Assignment 1).

1.2 Cleaning and Filtering

We start with a large amount hourly temperature data from the NOAA which included 300 global cities. We keep only the temperature data that passed all quality control checks and organize this data into 500 point time series where each measurement is exactly three hours apart. The data that is not within the Jan 1st, 1900 - Dec 31st, 2019 window is dropped only the data that corresponds with the calculated climate norms is analysed. This leaves us with 401845 analysable time series from 127 global cities.

1.3 Experimental groups

For this experiment, we use the classifier from Bury et al, 2021 which can identify the presence of early warning signals from Hopf, fold, and transcritical bifurcations from any selected point within a 500 point time series. Due to the flexibility of the classifier, we decide to create two main experimental groups. The first is bifurcation classified at point 400 and the second is bifurcation classified at point 450. After classification, the first group is left with 401845 analysable time series from 127 global cities while the second group is left with 372547 analysable time series from 116 global cities. This was due to us deciding to drop miss-classified data.

2 Model Assessment

2.1 ROC Curves

An ROC (Receiver Operating Characteristic) curve is a plot showing the performance of a classification model at all classification thresholds. On the y-axis of plot there is the amount of true positive events (correctly classified) and on the x-axis of the plot there is the amount of false positive events (incorrectly classified). Most ROC curves have there axis's ranging from zero to one to signify the percent of predictions on a zero to one scale that are correct and incorrect. The way that an ROC curve is generated is we calculate the model performance at many thresholds ranging from zero to one. At each calculated threshold, we plot a point on the ROC plot where the location of the point is dictated by the true positives and false positives dictated by the model at the specified threshold. When done with sufficient density, we can connect the points on the plot which generates our ROC curve.

A way that an ROC curve can be assessed is by calculating the ACU (area under the curve). The ACU measures the entire two dimensional area underneath the entire ROC curve. This can be calculated in a verity of way but is mostly done by numerical integration using the points used to make the curve. The ACU ranges from zero to one, where the model is 100% correct if the area under the curve is 1. If the area under the curve is zero then the model is 0% correct. If the model has an ACU > 0.9, it is considered to be strong. If the model has an ACU < 0.6, the model can be considered as weak.

2.2 Assessment Parameters

In addition to the ROC curves, we assess our model using the following parameters. We represent the number of true positive events as TP, the number of true negative events as TN, the number of false negatives events classified by the model as FN, and the number of false positive events classified by the model as FN.

- **Accuracy**: The accuracy of the model tells us the raw percentage of the classifications that the model correctly identities. $Acc \in [0, 1]$ and can be calculated by

  $$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

  where 0 is the lowest accuracy and 1 is the highest.

- **Precision**: The precision parameter identifies how well the model predicts the positive class. $Pre \in [0, 1]$ and can be calculated by

  $$Pre = \frac{TP}{TP + FP} \tag{2}$$

  where 0 is the lowest precision and 1 is the highest.

- **Recall (Sensitivity)**: The recall measures how well the model identifies true positive events. $Rec \in [0, 1]$ and can be calculated by

  $$Rec = \frac{TP}{TP + FN} \tag{3}$$

  where 0 means that the cannot identify true positive events and 1 means that the model correctly identifies all true positive events in the data.

- **F1-Score**: The F1-score is the weighted average score of the recall and precision. $F1 \in [0, 1]$ and can be calculated by

  $$F1 = \frac{2 * Pre * Rec}{Pre + Rec} \tag{4}$$

  where 1 is the best performance and 0 is the worse.

- **Specificity**: Specificity measures the rate of true negatives that are identified correctly. $Spc \in [0, 1]$ and can be calculated by

  $$Spc = \frac{TN}{FP + TN} \tag{5}$$

  where 0 means that the model cannot identify true negatives in the data and 1 means that the model correctly identifies all true negatives in the data.

- **Prevalence**: Measures how often positive events occur in the data. $Prv \in [0, 1]$ and can be calculated by

  $$Prv = \frac{TP + FN}{TP + TN + FP + FN} \tag{6}$$

  where 0 means that the data has almost no positive events while 1 means that the data has an extremely high amount of positive events.

3 Results and Discussion

We conduct the experiment with our two bifurcation groups and two subgroups (four total). Since we believe that Hopf bifurcations do not accurately represent heat waves, we decided to analyse our groups with and without the Hopf bifurcation classifier, creating the two subgroups for each full group. We assess the model at thresholds 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, and 1.00 for all of our groups.

3.1 Model Accuracy

We assess the model accuracy for our four groups at thresholds 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, and 1.00 using (1) (Table 1). There is an increase in model accuracy for each group as we increase threshold (Figure 1). We also see that the groups without Hopf are significantly more accurate then the groups with Hopf (Figure 1). The bifurcation at 450 group and the bifurcation at 400 groups have very similar accuracy when comparing there with and without Hopf groups (Figure 1). This suggest that a Hopf bifurcation does not represent a heatwaves or cold snap which contributes to low model accuracy. It also suggest that there is some other temperature phenomena(s) that are being identified by the Hopf classifier in our data which suggest some other temperature phenomena can be accurately represented by a Hopf bifurcation.

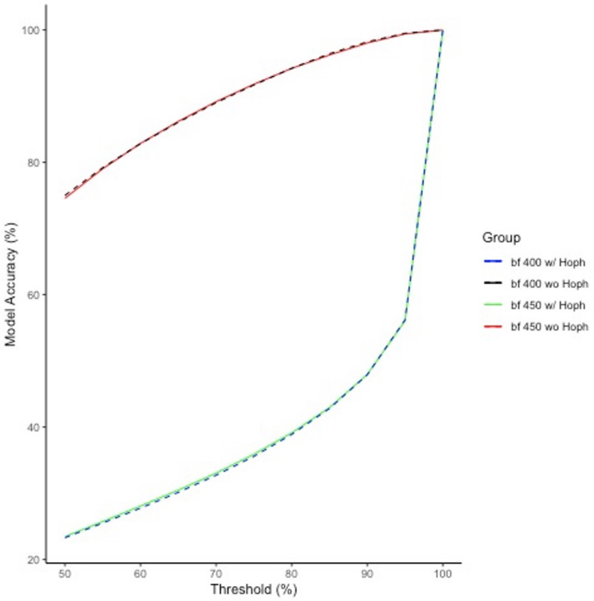


Figure 1: Model Accuracy

”Row”	”Thresholds”	”400_w_h”	”400_wo_h”	”450_w_h”	”450_wo_h”
”1”	0.5	0.233	0.75	0.234	0.746
”2”	0.55	0.255	0.792	0.257	0.79
”3”	0.6	0.278	0.828	0.281	0.828
”4”	0.65	0.301	0.86	0.305	0.862
”5”	0.7	0.327	0.89	0.33	0.892
”6”	0.75	0.356	0.917	0.359	0.918
”7”	0.8	0.389	0.942	0.391	0.942
”8”	0.85	0.428	0.964	0.429	0.962
”9”	0.9	0.479	0.982	0.48	0.98
”10”	0.95	0.562	0.995	0.56	0.994
”11”	1	1	1	1	1

Table 1: Model Accuracy

3.2 F1 Score

We calculate the F1 score for each group at each one of our selected thresholds. We see that our overall F1 scores are very low with the maximum score being 0.94 and the lowest being 0.51 (Table 2). The NA values in the table are due to the NA values for the precision at the 100% threshold (will be discussion in the next section). We observe that the F1 score peak at our lowest calculated threshold of 0.5. This suggest that our model either has a very low precision and threshold, or precision or threshold is significantly higher then its compliment in the formula (4). In our words, we observe that our model is not identifying true positives, true negatives, or both with a high degree of accuracy which will be identified in a later section.

"Row"	"Thresholds"	"400_w_h"	"400_wo_h"	"450_w_h"	"450_wo_h"
"1"	0.5	0.093	0.077	0.095	0.086
"2"	0.55	0.093	0.074	0.095	0.086
"3"	0.6	0.092	0.073	0.094	0.085
"4"	0.65	0.092	0.072	0.094	0.086
"5"	0.7	0.092	0.067	0.094	0.086
"6"	0.75	0.091	0.067	0.093	0.082
"7"	0.8	0.091	0.066	0.092	0.076
"8"	0.85	0.091	0.065	0.091	0.06
"9"	0.9	0.09	0.058	0.091	0.056
"10"	0.95	0.088	0.051	0.092	0.09
"11"	1	NA	NA	NA	NA

Table 2: F1 scores

3.3 Precision

We calculate the model precision for each group at each of our selected thresholds. Overall, we observe that our model has a very low precision for all of our groups which ranges between 0.027 and 0.05 (Table 3). The NAs are caused by a zero in the denominator in the formula at threshold one which will need to be further investigated before we get our next set of results for the project. We observe that the precision of the model is slightly higher at lower thresholds with the max calculated value being at a threshold of 0.5, 0.55, and 0.6 for most groups (Table 3). This suggest that our model is very bad at identifying only true positives in the data, either by not being able to identify true positives at all or identifying a bunch of false positive events as well. This implies that there are other temperature spikes and drops in the data that can be represented at our three bifurcations in addition to just heatwaves and colds snaps. We will need assess the model recall (next section) to see if our model is actually identifying heatwaves and cold snaps in the data as well, or if it is solely identifying other temperature events.

”Row”	”Thresholds”	”400_w_h”	”400_wo_h”	”450_w_h”	”450_wo_h”
”1”	0.5	0.049	0.04	0.05	0.045
”2”	0.55	0.049	0.038	0.05	0.045
”3”	0.6	0.048	0.038	0.049	0.044
”4”	0.65	0.048	0.037	0.049	0.045
”5”	0.7	0.048	0.035	0.049	0.045
”6”	0.75	0.048	0.034	0.049	0.043
”7”	0.8	0.048	0.034	0.048	0.04
”8”	0.85	0.047	0.033	0.048	0.031
”9”	0.9	0.047	0.03	0.048	0.029
”10”	0.95	0.046	0.027	0.048	0.049
”11”	1	NA	NA	NA	NA

Table 3: Precision

3.4 Recall

We calculate the model recall for each of our groups at each of our selected thresholds. Our model has very high recall of each of our groups excluding ”450 wo h” for all thresholds excluding 1 (Table 4). For ”450 wo h”, the model maintains a recall above 90% up to a threshold of 0.8 (Table 4). This suggest that our of our model is extremely good up to a threshold of 0.8 at identifying true positives in the data, which means that the model misses very little heatwaves and cold snaps at these thresholds. We observe that the model overall has a higher recall for the groups including the Hopf bifurcations (Table 4). This implies that a small amount of heatwaves and cold snaps are actually identified by the Hopf classifier so without it, the model is prone to missing some of these events. We also observe that the recall decreases as we raise the threshold (Table 4). The cause of this is all heatwaves and cold snaps do not cause the model probability to be above 0.8 which leads to an increase in false negatives in these higher thresholds groups and therefore a lower recall.

”Row”	”Thresholds”	”400_w_h”	”400_wo_h”	”450_w_h”	”450_wo_h”
”1”	0.5	0.999	0.985	0.998	0.981
”2”	0.55	0.999	0.982	0.998	0.975
”3”	0.6	0.999	0.979	0.998	0.97
”4”	0.65	0.999	0.972	0.998	0.966
”5”	0.7	0.999	0.965	0.998	0.955
”6”	0.75	0.998	0.954	0.997	0.94
”7”	0.8	0.998	0.929	0.997	0.907
”8”	0.85	0.998	0.892	0.996	0.83
”9”	0.9	0.997	0.792	0.995	0.692
”10”	0.95	0.996	0.484	0.994	0.536
”11”	1	0	0	0	0

Table 4: Recall

3.5 Specificity

We calculate the model specificity for each of our groups for each of our selected thresholds. The model specificity increases as we increase model threshold, with the groups without the Hopf classifier having a significantly higher specificity then the groups with the Hopf classifier (Table 5). This suggest that many of the false positives are caused by the Hopf bifurcation classifier which leads to the model having higher accuracy when the Hopf classifier is not included which has been shown in the general accuracy calculation. The increase in specificity with the increase in model threshold suggest that the higher the threshold, the more likely the model is more likely to properly identify a heatwaves due to the decease in false positives identified by the model (5).

”Row”	”Thresholds”	”400_w_h”	”400_wo_h”	”450_w_h”	”450_wo_h”
”1”	0.5	0.201	0.747	0.202	0.743
”2”	0.55	0.226	0.79	0.227	0.788
”3”	0.6	0.25	0.827	0.253	0.827
”4”	0.65	0.276	0.86	0.279	0.861
”5”	0.7	0.304	0.89	0.306	0.891
”6”	0.75	0.334	0.917	0.337	0.918
”7”	0.8	0.37	0.942	0.372	0.942
”8”	0.85	0.411	0.964	0.413	0.962
”9”	0.9	0.465	0.982	0.466	0.98
”10”	0.95	0.552	0.995	0.551	0.994
”11”	1	1	1	1	1

Table 5: Specificity

3.6 Prevalence

We calculate the prevalence for all of our group at all selected thresholds. We observe that the prevalence is very low for all groups, with prevalence decreasing as we increase thresholds (Table 6). This suggest that that there is a very low abundance of heatwaves and cold snaps in the data which is expected due to the definition of a heatwave and cold snap (see Assignment 1). It also suggest that there is a decrease in events as we raise model threshold which is also expected as the model is less likely to identify positive events if the threshold is higher. Prevalence between all groups is very (Table 5).

”Row”	”Thresholds”	”400_w_h”	”400_wo_h”	”450_w_h”	”450_wo_h”
”1”	0.5	0.039	0.011	0.04	0.012
”2”	0.55	0.038	0.008	0.039	0.01
”3”	0.6	0.037	0.007	0.037	0.008
”4”	0.65	0.035	0.006	0.036	0.007
”5”	0.7	0.034	0.004	0.035	0.005
”6”	0.75	0.032	0.003	0.033	0.004
”7”	0.8	0.031	0.002	0.031	0.003
”8”	0.85	0.029	0.001	0.029	0.001
”9”	0.9	0.026	0.001	0.026	0.001
”10”	0.95	0.021	0	0.022	0.001
”11”	1	0	0	0	0

Table 6: Prevalence

3.7 AUC

We generate ROC curves for each on of our thresholds (Figure 3 as example) and calculate the area under each ROC curve for each of our thresholds (Figure 2). We see that for the bifurcation at 450 group, the area under the curve ranges from 0.5335 to 0.5463 and is the same for both with and without Hopf (Figure 2). The bifurcation at 400 group, we see that both groups have different are under there curves with the without Hopf group having a slightly higher accuracy then the with Hopf group for all thresholds (Figure 2). For both groups, the area under the curve is constant for all thresholds (Figure 2). This suggest that our model has a low precision at identifying true positive events which is verified by our precision calculations (Section 3.3). This means that even tough our model is quite successful at identifying heatwaves and colds snaps in the data, the model is also picking up other events in the data which means we need to refine our method and our definition to accurately identify these events in the data.

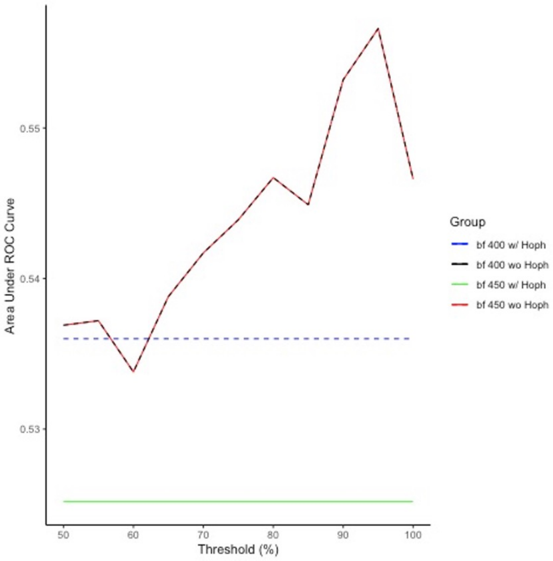
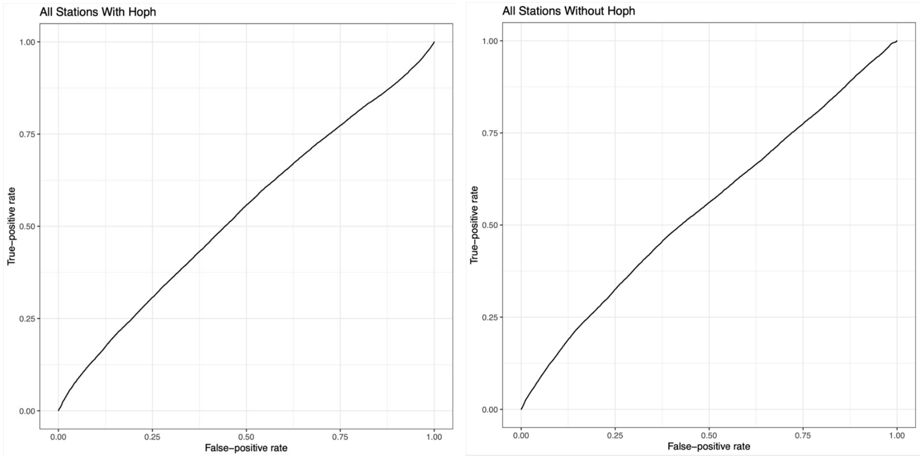


Figure 2: AUC



(a)

(b)

Figure 3: ROC Curves

4 Future Research

I believe our results are positive as they show us that our theory is not wrong which signified by the high model specificity and recall for our without Hopf groups. Overall, these results serve as a really strong stepping stone to gain our intended results. This opens the door for future research to refine our method, improve our model accuracy, and gain our intended results. This includes

- Testing the model on the bifurcation at point 500 group
- Analysing all three bifurcation classifiers separately to identify and remedy issues
- Analyse current results to see what is causing all of the false positives, especially for the Hopf classifier
- Analyse a multidimensional time series including other parameters such at pressure, wind speed, wind directions, ect. This is because anomalies in other weather parameters are know to cause heatwaves and cold snaps which means that there may be EWS in the other parameters as as well leading to heatwaves and cold snaps.
- Revisit and modify current definitions of heatwaves and cold snaps. This may lead to an increase accuracy in our model.
- Lag window analysis. This means analysing up to a 100 point period after the bifurcation point to see if there is a specific time that the heatwaves and cold snaps occur after the model predicts the bifurcation.

References

[1] Bury, T.M., Sujith R. I., Pavithran I., Scheffer M., Lenton T.M., Anand M., Bauch C.T. (2021). *Deep learning for early warning signals of tipping points.* Proceedings of the National Academy of Sciences of the United States of America. 118(39).

[2] DataTechNotes. (2019). *Precision, Recall, Specificity, Prevalence, Kappa, F1-score check with R.* DataTechNotes A blog about data science and machine learning https://www.datatechnotes.com/2019/02/accuracy-metrics-in-classification.html

[3] Google For Developers. (2022). *Classification: ROC Curves and AUC.* Google https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[4] Wang, C., Wang, Z.-H., Sun, L. (2020). *Early-warning signals for critical temperature transitions.* Geophysical Research Letters, 47