Characterizing West-Nile Virus Dynamics with Explainable Machine Learning

Ishaan M. Sinha *
South Brunswick High School
Monmouth Junction, NJ

Nikhil Muralidhar
Dept. of Computer Science
Stevens Institute of Technology
Hoboken, NJ

---

[*] All images original, created by this author

Abstract: The West Nile Virus(WNV) is a seasonal arbovirus (anthropod-borne-virus) spread through avian hosts and mosquito vectors. The WNV has emerged as the leading cause of mosquito borne disease in the continental US, and poses a significant burden on our society. Hence, monitoring the evolution of WNV is imperative for controlling and possibly preventing the spread of the virus to large populations. However, due to the lack of extensive data on the spatio-temporal dynamics of mosquito and bird populations, employing traditional mechanistic models for estimating the seasonal evolution of this virus is infeasible. Hence, in this work, we employ sophisticated data-driven (specifically, deep-learning) models in conjunction with a low volume of historical WNV, arboviral and climate data, and present a state-wide analysis of WNV dynamics over the entire continental US. In our work, we develop an *explainable* deep learning solution to forecast the short-term (1-month ahead) and long-term (6-month ahead) spread of WNV in each US state and via rigorous experimentation, report good prediction results compared to popular time-series forecasting models as well as models designed specifically for disease forecasting. Additionally, we analyze the evolution of WNV by identifying important drivers for disease spread, and forecasting future evolutionary dynamics by employing an explainable deep learning model. Specifically, we adapt a state-of-the-art deep learning transformer model to understand the effects of eco-climactic and arboviral indicators on the spread of WNV, across different states in the US. Through rigorous analysis, we demonstrate that our model is robust and accurately predicts the start of the seasonal dynamics (imperative for preemptive planning), national trends, temporal trends of the disease. Moving forward, our work will serve as a solid foundation, for developing an interpolating disease operator which can predict disease cases at finer temporal and spatial granularities (i.e., county-level, city-level, weekly time-scale).

Keywords: Disease Modeling, Time Series Forecasting, West Nile Virus, Deep Learning, Explainable AI.

1 Introduction

West Nile Virus(WNV), of the Flaviviridae family and Flavivirus genus, came to the United States in the summer of 1999 and has established itself as the leading cause of domestically acquired arboviral(anthropod borne virus) disease in the United States [9] [20]. Following its massive expansion in North America in 2003[15] , the virus settled into a cycle between avian reservoir hosts and mosquito vectors (See figure 1)[10]. Humans serve as dead-end hosts for the virus. For humans, it is estimated that $20 - 30\%$ of WNV infections result in symptoms, and 1% result in neuroinvasive disease [18]. It was not until 2013, when human infection levels reached close to the initial peak in 2003 (See figure 2), that scientists realized the virus will continue to unpredictably outbreak regionally throughout the United States[8]. While the virus strikes annually, currently, we only have limited methods to forecast the magnitude and length of the annual outbreak.

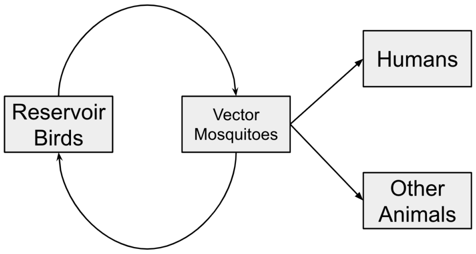


Figure 1: WNV Transmission Cycle. Humans become accidental dead end hosts.

We analyze the spread of WNV spatially (See figure 3). The first major outbreak in 2003 was mostly contained in Texas and Colorado. Since then, California has become the harshest breeding ground of the virus, primarily due to its mosquito-prone heat and dense urban cities. Furthermore, the Northeast has seen a recent development in cases, with New York and Pennsylvania populations starting to harbor great numbers starting 2018.

The states with the highest WNV count cases since the virus came to North America are California, Texas, and Arizona, making up 36.9% of US West Nile Virus cases from all collected years. Immediately, we notice that Arizona has an inexplicable meteoric rise in cases in 2022 (See figure 4). This incident has been well studied by the Center for Disease Control and cannot be explained by the data. It is not possible for any disease forecasting

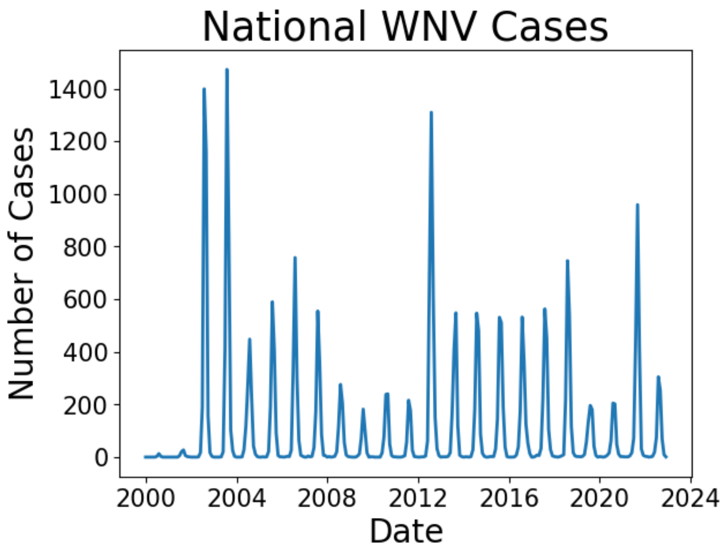
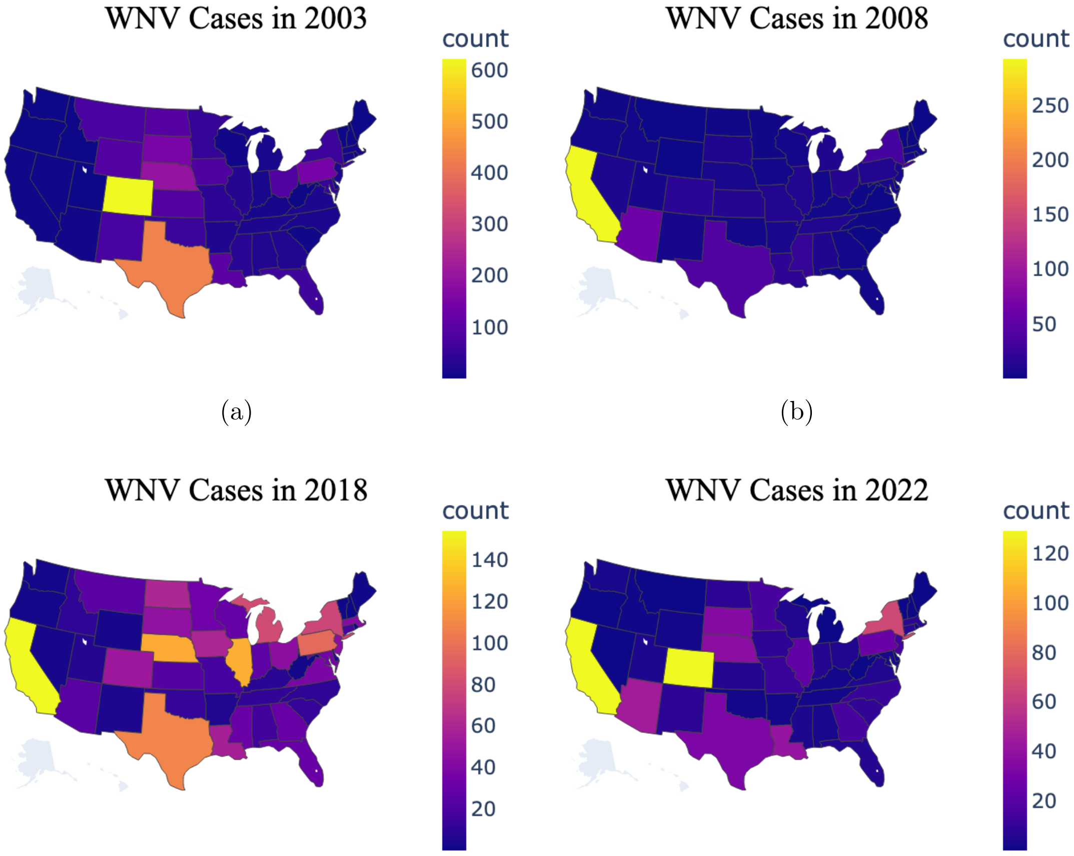


Figure 2: West Nile Virus human case counts in the United States from 2000-2022 at monthly intervals.



(a)

(b)

(c)

(d)

Figure 3: Full year WNV case count heat maps across the United States

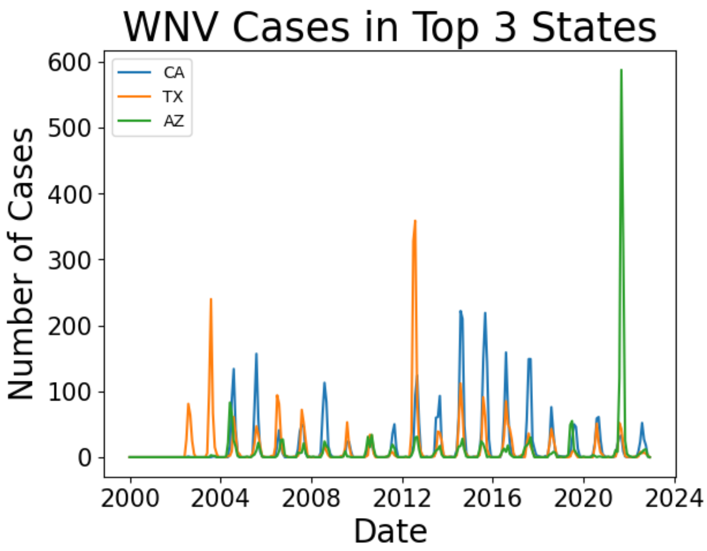


Figure 4: Monthly West Nile Virus cases in the three most prone states: California, Texas, Arizona

model to predict that rise.

Currently, there is no vaccine or treatment for WNV, so protection primarily relies on preventative actions (ex. mosquito repellent) [7]. If outbreaks can reliably be predicted several months in advance, public health efforts can more reliably coordinate and use limited resources to prevent the massive spread.

Furthermore, even after an outbreak begins, it has remained difficult to predict the future characteristics of the epidemic curve, due to the nonlinearity of the differential equations describing the West Nile Virus[3].

We aim to build a deeper understanding of the dynamics in different regions, ie statewide. In particular, we want to answer the following questions:

- How much emphasis does the spread of West Nile Virus in different states put on different categories of data: weather, infected mosquito populations, or temporal inputs?
- For different states, does the transformer pay more attention to recent data, or more distant data?
- For different states, the loss of which data source will weaken the model the most? This is concurrently a measure of robustness.

The Center for Disease Control (CDC) conducts an annual competition to model this virus, providing additional motivation to understand the epidemiology at play. They provide monthly statewide infection counts as well as collected mosquito data curated by VectoSurv. [2] [13].

In our work, we use a recently developed transformer architecture, called the Temporal Fusion Transformer [16], to model the disease and provide explainability with regards to the most relevant data sources for differing states, along with the differing disease dynamics in different states. This will allow epidemiologists and other disease modelers to better understand the complexity of modeling the disease, as the spread of WNV in different environmental locations has unique transmission reasons.

2 Background

The main challenge of forecasting this virus is the data paucity with regards to bird and mosquito population. Trying to fit the differential equations without this data yields, using parameter estimation, bad results due to the instability of the equations. Current mechanistic compartmental models rely on these sparse data sources, resulting in lower model accuracy [5].

Let S be the population of susceptible humans, E be the asymptomatically infected human population, I be the population of symptomatically infected humans. Note that the CDC disease counts is the derivative of I with respect to time [1]. The model differential equation was presented in [3]

$$\frac{dS}{dt} = \Pi_h - \frac{b_2\beta_3 M_i S}{N_H} - \mu_H S$$

$$\frac{dE}{dt} = \frac{b_2\beta_3 M_i S}{N_H} - \mu_H E - \alpha E$$

$$\frac{dI}{dt} = \alpha E - \mu_H I - \delta I$$

Despite using several strong optimizers and boundary conditions, numerical solvers were unable to estimate good parameters to fit the actual data. So, we believe that mechanistic disease modelers are not as powerful in comparison to data-driven architectures when it comes to United States West Nile Virus Modeling.

Previous work has used epidemiological differential equations as a physics guided loss in a regression neural network [21] [24]. Other data-driven work has primarily used gradient boosting models to understand feature importance for WNV for Europe [9][22]. We extend to use transformer models [23, 16, 26] for finer locations, states, in the United States.

Current research considers several biotic and abiotic features to be of importance for WNV transmission [9]. Some biotic features include: temperatures, precipitation, vegetation, mosquito vector abundance. Some abiotic features include: economic conditions, demographic characteristics. We choose to not use abiotic features in our model due to the coarseness of the measured data. Since we want to make monthly predictions using the WNV data(which starts in 2000), and the US Census publishes these demographic data every ten years, it would not greatly impact the model.

The other hindrance with forecasting WNV is the delay in data. The CDC publishes full WNV case counts at the end of each calendar year, as collating the number of infections from hospitals across the country comes with delay [6]. The weather data and arboviral data are published monthly and biweekly respectively, so for our monthly forecasts, do not provide any delay.

For this paper, we primarily focus on disease forecasting in the ten states with the highest counts, which contribute to 66.1% of cases in the country. These are: California, Texas, Arizona, Illinois, Colorado, Louisiana, Michigan, Nebraska, Mississippi, New York.

3 Problem Formulation

We want our multi-horizon forecasts to be in accordance with a scenario of real life forecasting, as in the CDC West Nile Virus Forecasting competition.

The task of our model is as follows: For endogenous data, we have a series of monthly West Nile Virus counts $\{y_1, .., y_n\}$. We want to forecast disease for the next 6 months, $\{\hat{y}_{n+1}...\hat{y}_{n+6}\}$. Next, we move the model ahead temporally, with West Nile Virus counts $\{y_1, .., y_{n+1}\}$. This time, the model forecasts $\{\hat{y}_{n+2}...\hat{y}_{n+7}\}$. We move temporally ahead in this manner, with the last model forecasting $\{\hat{y}_{n+19}...\hat{y}_{n+24}\}$ using the data $\{y_1, .., y_{n+18}\}$. Notice that this formulation is in line with the CDC competition, which asks for 6 month ahead forecasts every month.

For the exogenous data, each y_k is associated with some number of covariates, say m covariates. We lag our data to ensure there are covariates for future predictions. In the following algorithm(See figure 1), we let $\mathcal{A}$ be the dataset, flen be the number of months we want to forecast with each model, and fhor be the total length of the forecast horizon. For the purposes of this paper, in guidance with the CDC, we forecast 6 months at a time, so flen = 6 and we forecast for the next 24 months, fhor = 24. But, the results are easily transferrable for different forecasting lengths or forecast horizons.

Algorithm 1 CDC WNV Forecasting Imitator

```
function FORECAST($\mathcal{A}$, flen, fhor)
    $N \leftarrow$ length(A)
    for $k \leftarrow N-$fhor to $N-$flen do
        data $\leftarrow$ first k rows of $\mathcal{A}$
        train TFT model $\mathcal{M}$ using data
        prediction $\leftarrow$ predictions from M
        save prediction
```

3.1 Transformer Architecture

We now describe the Temporal Fusion Transformer, which is composed of, among others, Gating mechanisms, Variable Selection networks, Static Covariate Encoders [16].

Of particular importance to us is the Variable Selection networks, which learns the importance weights of each input feature. This is important for interpretability, as well as reducing the impact of unnecessary noisy inputs.

Let E_t^j be the transformed (after a power transform) jth variable input at time t. Now, let $\mathcal{E} = [E_t^j, ..., E_t^m]$ where there are m variables. Now, the variable selection weights come by putting $\mathcal{E}_\sqcup$ through a Gated Residual Network, followed by a Softmax layer:

$$v = \texttt{Softmax}(GRN(E_t))$$

At each time step, each variable is processed through its own GRN. The overall weight for each feature is a summation of its individual weights for each time period.

The other vital mechanism of the TFT is its interpretable multi-head attention. This self attention mechanism learns long-term relationships in the data and is common within transformer based architectures. We aggregate the attention of each MultiHead layer to get attention weights for each time period.

We include a representation of the overall workflow (See figure 5), masking the details of the TFT.

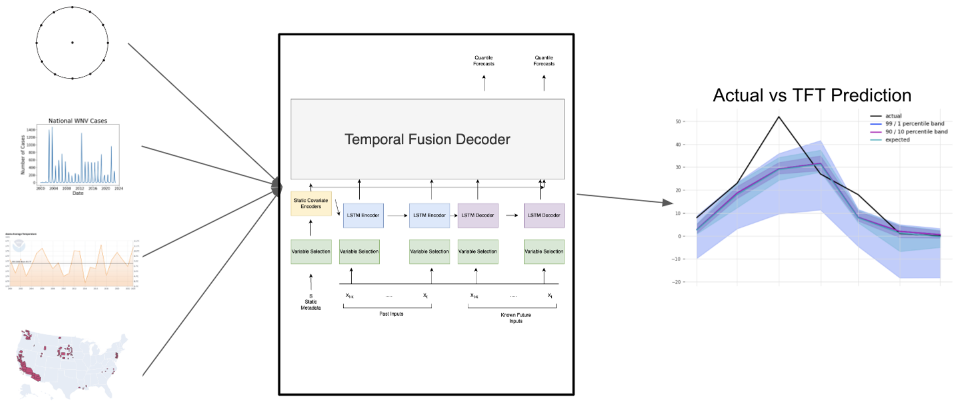
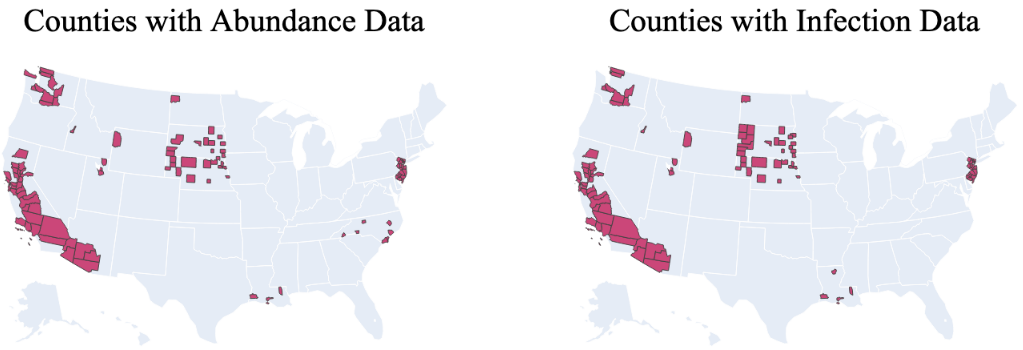


Figure 5: Pictorial representation of prediction pipeline

In the transformer models trained for this paper, we use a batch size of 32, dropout of 10%, 2 LSTM layers, and sequences of length 32 to predict the next 6 months.

4 Experiment Setup

We use our Temporal Fusion Tranformer model [11] to predict the West Nile Virus in the 10 states with the highest cases from the year 2000, of most importance to policy makers. These range from 790 cases in New York from the years 2000-2022 to 4496 cases in California in the same time range. Since our accepted CDC data for the West Nile Virus is fully accurate to December 2022, we treat the years 2021 and 2022 as our test set, along which we want to make predictions. To reenact a real forecasting challenge, we have trained a new model for every additional month of data, predicting the next 6 months. So, since we want to predict


Counties with Abundance Data

Counties with Infection Data

Figure 6: Notice the scarcity of mosquito data across different US regions

for the next 24 months, and in accordance to the CDC guidelines of wanting to forecast values for the next 6 months, we have 24 - 6 + 1 = 19 models per state. From each of these models, we collate the 1-month ahead, 2-month ahead,. . . , 6-month ahead predictions.

4.1 Dataset Description

We make use of five weather features from the National Oceanic and Atmospheric Administration: precipitation, max temperature, min temperature, average temperature, and z-index. The Palmer Z-index is a measure of short-term dryness [19]. Each of these features is taken at monthly granularity.

Next, we make use of arboviral data, combined by VectorSurv [2], to derive two important variables identified by CDC experts: the mosquitoes per trap night and the infection rate. This data was collected by setting traps across certain counties in the country and collection data from caught mosquitoes. The mosquitoes per trap night is a measure of the vector population. The infection rate is a measure of the percent of the vector population which is infected. This was more correlated with the WNV data than external sources, like Wikipedia page visits. [25]

For this collected mosquito data, the traps do not cover all regions of the United States(See figure 6). There is a heavy emphasis on collecting data from California and New Jersey, but data is sparse elsewhere. So, to provide insight to the transformer model, we take the maximum, minimum, and average for all collected counties at a given time. As a rough proxy, we use these as features for all states. We also use some endogenous data, namely the West Nile Virus case count one year prior to the prediction data along with the national count one year prior, in case the dynamics of the state follow the national count better. Lastly, we want to provide some temporal encoding to the model, which indicates the cyclic nature of the West Nile Virus. Hence, we use a cyclical encoding for the month of the year as input. This entails passing in the x and y coordinates for 12 equally spaced points of a circle, for each month of the year [27]. See figure 4.1

Feature Category	Feature Name	Feature Name in Model
Climate	Average Monthly Temperature	tavg
	Maximum Monthly Temperature	tmax
	Minimum Monthly Temperature	tmin
	Monthly Precipitation	pcp
	Palmer Z-Index	zndx
Arboviral	Average Mosquito Infection Rate	infectionRate
	Maximum Mosquito Infection Rate	infectionRate.1
	Minimum Mosquito Infection Rate	infectionRate.2
	Average Mosquitoes Caught per Trap Night	rt1
	Maximum Mosquitoes Caught per Trap Night	rt2
	Minimum Mosquitoes Caught per Trap Night	rt3
Endogenous	Prior year Mosquito Count in state	count_yearprior
	Prior year national Count	national_count_yearprior
	Six month prior national Count	national_count
Temporal Encoding	$\cos n/12$ where n is the month number from 0-11	month_cos
	$\sin n/12$ where n is the month number from 0-11	month_sin

4.2 Baselines.

To ensure that the transformer model outperforms standard diseases modeling architectures, we used the same data to train other models.

- ARIMAX [4]: This is a well studied standard linear time series model. We optimize the parameters using a grid search for the optimal lag and optimal moving average component.

- LSTM [12]: This is a well studied recurrent neural network architecture used in deep learning. It can keep track of long term dependencies and has been widely used for time series forecasting. The model we trained uses 16 length inputs to predict 1 output at a time. We use a learning rate of 0.1%.

- Zero-Inflated Poisson [14]: Zero-Inflated models are commonly used for count data, with frequent zero-valued observations, as with the WNV case count. The distribution of counts is modeled as a Poisson distribution with probability mass function:

  $$f(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

  where λ is a fittd parameter and k is the number of cases.

5 Results & Discussion

5.1 Model Mean Absolute Error + Predictions

To quantitatively understand the strength of each trained model, we use the mean absolute error:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

where y_i is the true disease count and x_i is the predicted disease count.

For 7/10 of the states, our TFT model outperforms all the baselines for the 1 month ahead predictions (See figure 1). Furthermore, the TFT model provides a 54% performance boost over the ARIMAX, 15% performance boost over the LSTM, and a 7.2% performance boost over the Zero Inflated Poisson model.

Similarly, our model outperforms the baselines for 7/10 states in the 6 month ahead predictions. The TFT model provides a 51% performance boost over ARIMAX, 26% performance boost over the LSTM, and an 86% boost over the Zero Inflated Poisson

State	1 Month Ahead				6 Month Ahead			
	ARIMAX	LSTM	ZIP	**TFT**	ARIMAX	LSTM	ZIP	**TFT**
CA	7.46	6.35	5.7	**4.12**	14.45	14.03	**8.08**	10.08
TX	10.16	3.70	5.73	**3.66**	17.3	15.36	8.67	**6.11**
AZ	127	58.76	57.12	**56.6**	144.22	67.23	712.14	**58.93**
IL	5.21	4.64	3.49	**2.04**	6.62	7.54	4.39	**3.8**
CO	9.16	5.2	**3.9**	4.87	13.44	10.79	**8.8**	10.65
LA	2.27	2.13	1.58	**0.9**	4.12	2.61	2.66	**1.84**
MI	3.4	4.58	1.64	**1.36**	4.82	8.6	**3.01**	3.51
NE	4.67	3.75	**2.86**	3.71	**4.55**	8.79	6.61	6.07
MS	1.78	1.48	1.76	**0.36**	3.32	2.22	2.19	**1.24**
NY	1.87	2.29	**1.23**	1.29	4.01	4.59	3.85	**3.12**
With AZ Avg.	17.3	9.29	8.50	**7.89**	21.7	14.2	76.0	**10.5**
Without AZ Avg.	5.11	3.79	3.10	**2.48**	8.07	8.28	5.36	**5.16**

Table 1: Collated errors for 1 Month Ahead and 6 Months Ahead predictions for each state

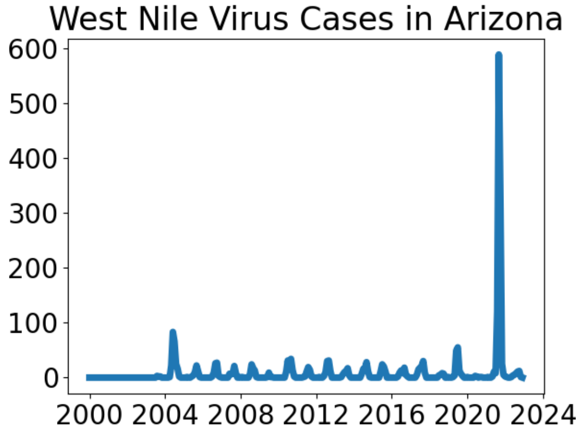


Figure 7: Recent Anomalous WNV Dynamics in AZ

As predicted, none of the models are able to predict the Arizona case counts. This was the largest ever outbreak in a US state, over four times greater than the last largest outbreak in Arizona, in 2004. None of the climatic or collected mosquito data gives reason to this, hence the models are unable to predict it.

Looking at the 1 month ahead Mississippi predictions(See figure 9), it is notable that the Zero-Inflated Poisson, LSTM, ARIMAX models tend to greatly overpredict the case count. The TFT model accurately predicts the first nonzero months, signifying the start of the

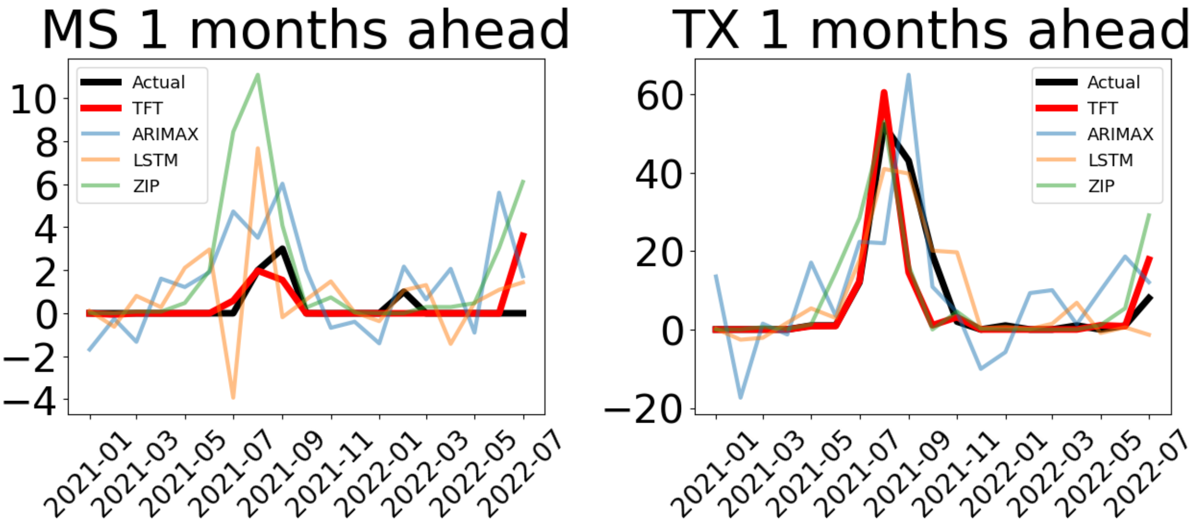


Figure 8: 1 Month Ahead Predictions for Mississippi and Texas, with the TFT model and ground truth values bolded

WNV season. The tansformer model also learns to not predict negative case counts, without writing in these explicit bounds in the architecture.

For the Texas 1 month ahead predictions, the TFT accurately predicts the peak in the season as well as temporal regions with no WNV cases. Once again, the transformer picks up on the starting of the West Nile Virus season much more effectively than the other baselines.

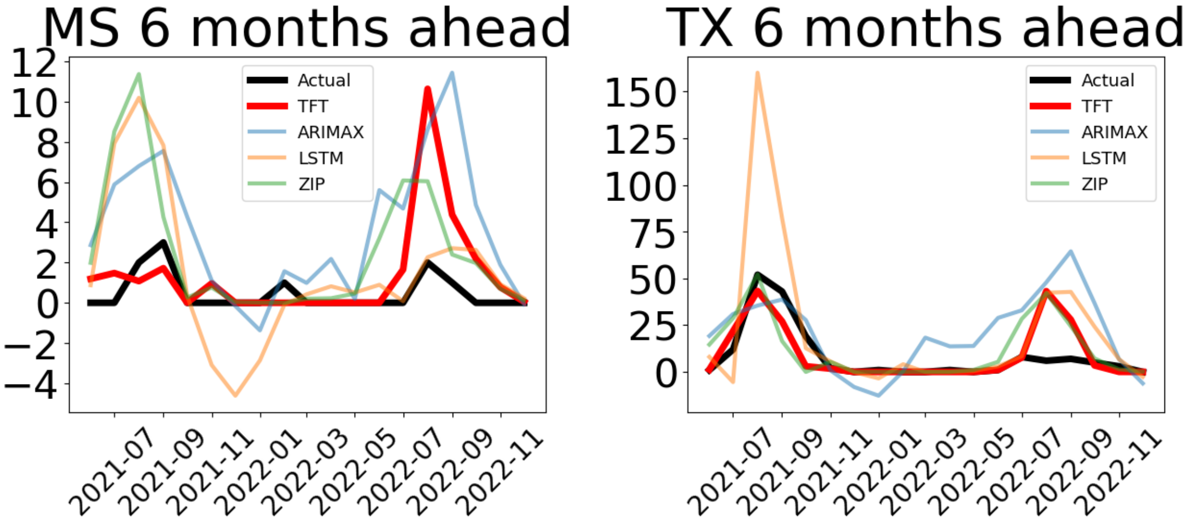


Figure 9: 6 Month Ahead Predictions for Mississippi and Texas, with the TFT model and ground truth values bolded

As expected, all the models perform worse for the 6 month ahead forecast. Once again, most models seem to be overpredicting the case counts. But, in Texas, the TFT model is able to predict the first peak accurately, even 6 months in advance. In general, the TFT has learned about the start of the season more effectively than the other models. But, all models give a false positive during August, 2022 in Texas.

Overall, our TFT model is far superior than the baselines, so we look into its mechanisms to further understand the feature importance for WNV spread.

5.2 Feature Importance

We want to analyze the importance of different features in different states, with regards to the encoder of the transformer. The encoder provides a vector representation of the inputted data, so can give vast amounts of information with regards to the features it gives higher weight.

To visualize, for each state, across all the models for that state, we average the importance given to each of the 15 features. Then, we curate the top 10 features. To analyze differences between the features at a glance, we compute the Jaccard Similarity between these sets of 10 features (See figure 10), letting A,B be sets, is:

$$\texttt{Jaccard Similarity} = \frac{|A \cap B|}{|A \cup B|}$$

Immediately, it is apparent that Arizona differs from all the other states, likely due to the anomalous case counts in 2021. Also, Nebraska, Texas, and Colorado all put importance on the same top 10 features. Their geographic closeness corroborates this. Mississippi and Louisiana are neighboring states, so it makes sense that the WNV dynamics in both of those states focus on the same features.

In California (See figure 11), there is great importance given to the maximum mosquito infection rate and the z index. The z-index is a measure of dryness/wetness which serves as a proxy for the NDVI vegetation index for the region. Since humidity often controls the mosquito population in the area, it is apparent that the West Nile Virus dynamics in California are heavily dependent on the infectious mosquito population. Other factors, such as the precipiation, are less relevant.

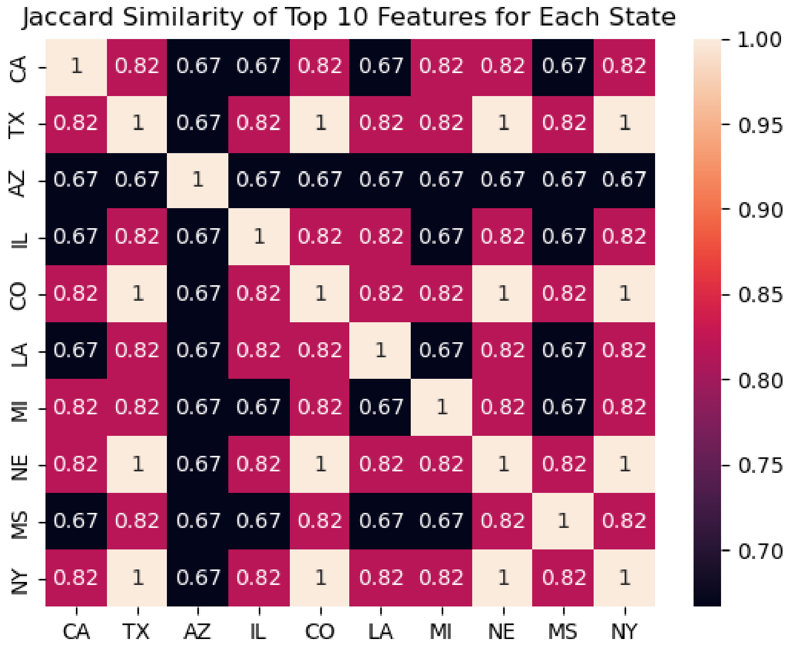


Figure 10: Heatmap of Jaccard Similarity between Top Features

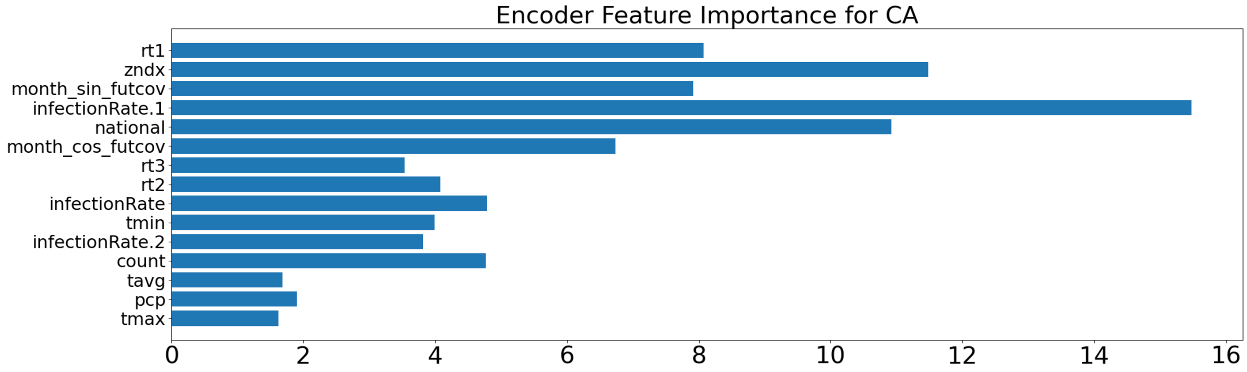


Figure 11: Encoder Feature Importance in California

In Arizona (See figure 12), there is great importance on the collected arbovirus maximum mosquito infection rate, in addition to the monthly temporal encoding, specifically the month sin. Just as in California, we notice that infected mosquito population is strongly relevant, as infected mosquito bites transfers the virus to humans. Of other importance is the dependency on time of year. This suggests that the WNV dynamics in Arizona more closely follows the time of year, as opposed to, say weather variables.

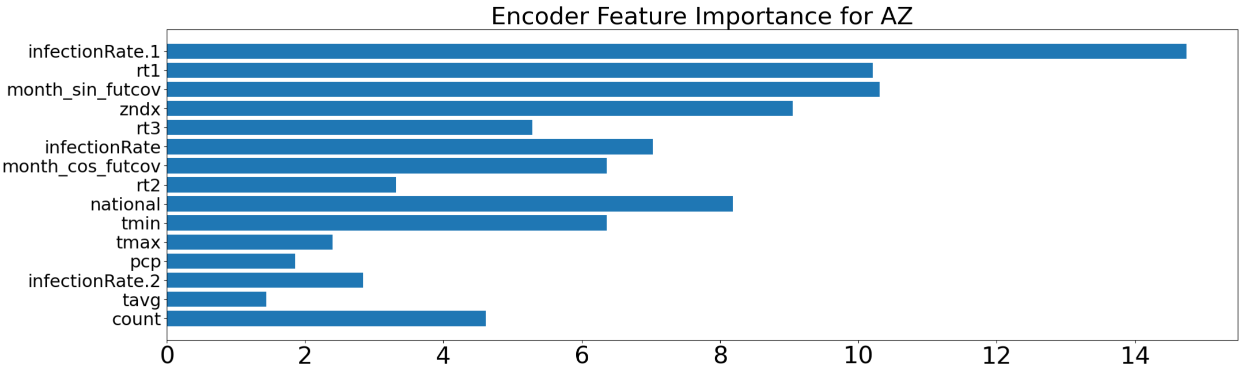


Figure 12: Encoder Feature Importance in Arizona

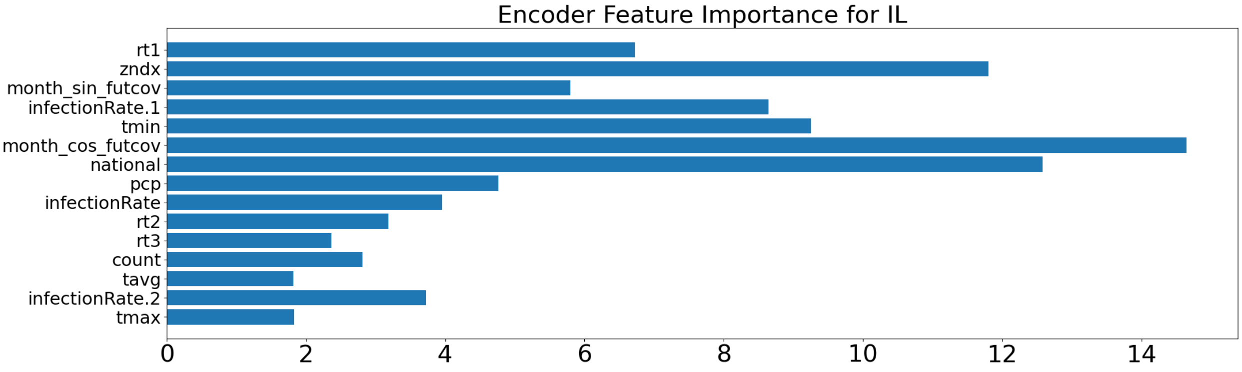


Figure 13: Encoder Feature Importance in Illinois

In Illinois (See figure 13), there is great importance on the temporal encoding, along with the national case count. So, just as in Arizona, the WNV dynamics in Illinois heavily depend on the time of year. Interestingly, the importance on national count suggests that the dynamics in Illinois follow the national WNV trend. So, Illinois seems to not rely much on local data and dynamics.

Averaging throughout all the states, the arbovirus data and the temporal encoding have emerged as the most useful features. Since the WNV season usually occurs from June to November, and the infection rate and populations of mosquitoes increase the spread of disease, this makes sense. Overall, by analyzing the different feature importances for different states, we understand the differing WNV dynamics present. This corroborates the difficulty of creating a model that accurately predicts WNV for all states, because different

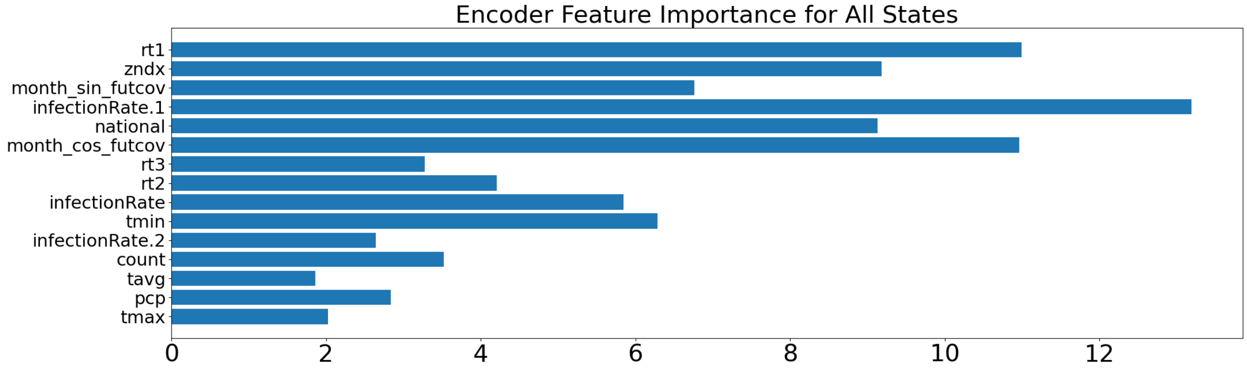


Figure 14: Encoder Feature Importance averaged for All States

geographical locations have WNV transmission dependence on different features.

5.3 Attention

The most important part of every transformer architecture is the attention, namely the ability to focus on certain data inputs, and put less emphasis on others. We analyze which temporal inputs the TFT puts greatest emphasis on.

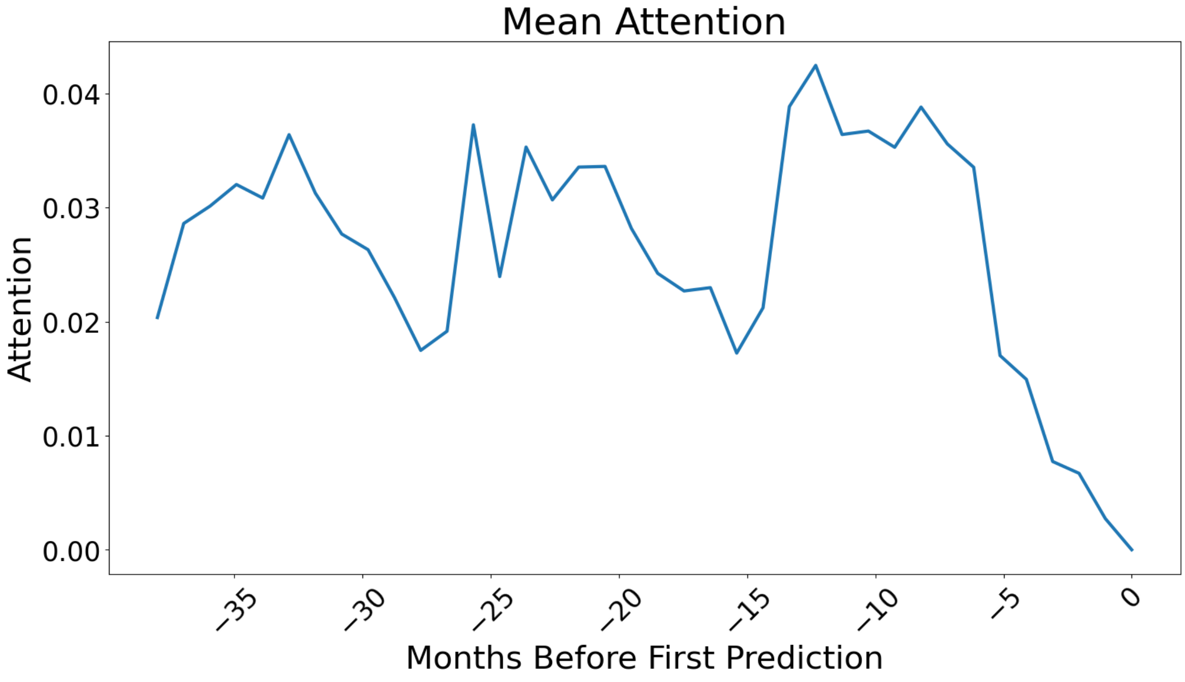


Figure 15: The transformer attention for each month prior to the first forecast

Looking across all states(See figure 15), we notice that the models pay most attention to roughly 12, 24, 36 months prior to each forecast. This make contextual sense as the number of WNV cases in month m of year n is strongly correlated to the cases in the same month m of years $n-1, n-2, n-3$. Of surprise is that there is not much attention on the most recent data, suggesting that the WNV dynamics are not very dependent on the most recent data inputs, including the most recent amount of cases.

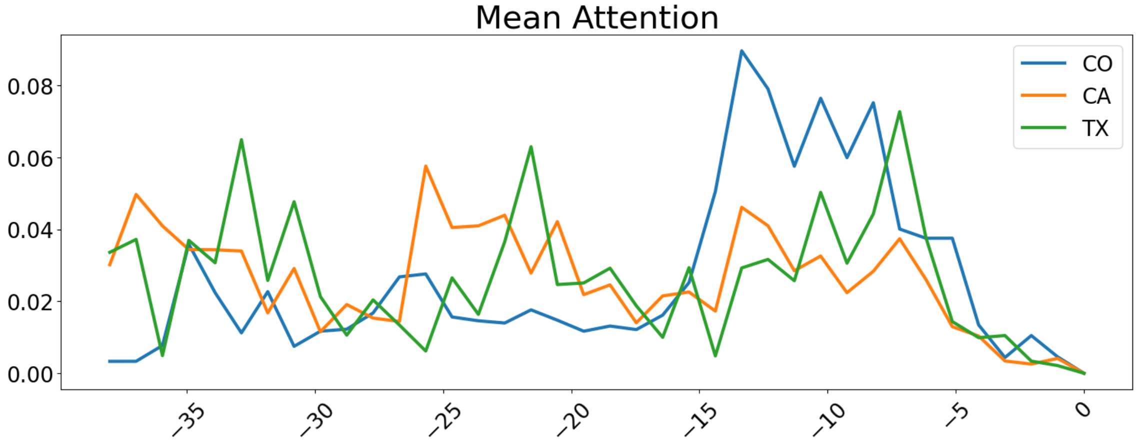


Figure 16: Mean Attention for Colorado, California, Texas

Looking deeper at individual states gives us a better idea of the different dynamics at play in different states (See Figure 16).

For example, in Colorado, there is greater attention given to the more recent data, implying that the West Nile Virus dynamics are more strongly correlated to recent trends in the data.

In difference, California seems to give roughly the same attention to each of the last three years of data, meaning that there is not a greater emphasis on recent dynamics, rather those on a longer range scale.

This is partly the reason for the challenge of modeling the West Nile Virus. Not only is there a data scarcity in relation to the bird and mosquito population data, which are the hosts and transmitters of the virus, but different ecoclimactic regions follow different trends with the data.

5.4 Ablation Analysis

We perform an Ablation analysis on our models, where we ablate each covariate, by removing that specific exogenous input from the training of our models, and analyze the mean absolute errors per state. This gives us greater insight into the most important features for each state.

Feature	CA	TX	AZ	IL	CO	LA	MI	NE	MS	NY
count yearprior	**13.91**	6.54	58.97	**7.97**	7.87	1.83	2.26	5.19	1.06	3.01
infectionRate	11.84	6.83	57.51	3.26	13.88	1.39	2.88	4.45	1.33	4.25
infectionRate.1	9.69	6.7	58.06	3.53	12.29	1.36	2.45	5.81	1.14	2.79
infectionRate.2	8.42	**8.37**	57.13	3.84	7.14	1.59	2.21	4.3	1.05	3.22
month cos	11.85	7.27	57.25	4.33	7.07	**2.05**	1.71	4.02	0.51	3.58
month sin	7.44	5.1	**59.27**	3.44	7.42	1.79	2.19	3.76	0.78	2.99
national count	7.62	6.72	56.43	4.15	11.93	1.25	2.43	4.53	1.07	2.97
nationalcount yearprior	6.59	7.19	57.48	3.78	6.87	1.67	2.03	4.06	0.44	2.78
pcp	7.48	5.71	58.86	3.43	13.28	1.37	2.87	8.24	0.55	2.45
rt1	8.62	5.53	58.43	4.18	6.44	1.36	2.41	4.2	1.2	2.14
rt2	11.19	7.15	57.67	4.28	10.27	1.6	**4.05**	5.35	1.01	3.23
rt3	9.34	7.43	57.52	4.8	6.63	1.09	2.08	**8.25**	**1.5**	3.81
tavg	7.7	6.55	58.04	5.02	14.43	1.49	2.6	6.76	0.77	2.95
tmax	7.04	6.97	57.61	3.39	7.61	1.37	2.43	4.55	0.81	2.64
tmin	7.18	7.36	58.46	6.82	**31.7**	1.48	2.25	6.21	0.75	**6.14**
zndx	6.51	7.91	58.44	5.62	12.73	1.28	1.81	6.28	0.71	5.69

Table 2: Ablated errors with the leftmost column feature removed from traning the model

Looking at the feature who's loss impacts the model's performance the most (See figure 2), the maximal mode is 2 for a singular feature: count yearprior, rt3, tmin – an endogenous source, an arbovirus source, a weather variable. The variability in the types of data most pertinent to the statewide models further highlights the different WNV dynamics in different states.

Furthermore, the models are also fairly robust, the loss of a singular data source does not impact the performance too greatly.

6 Conclusion & Future Direction

The analysis of the Temporal Fusion Transformer, especially with its feature importance and attention, has given great insight into the differences between WNV transmission in different states of the United States.

The TFT model is far superior for disease forecasting in comparison to other baselines, especially in the time frame the CDC requires for their competition.

Different states have different transmission dynamics. As an example, the California spread mostly relies on mosquito data and abundance. In contrast, Illinois follows more closely with the time of year and the national count. This means that the data inputs for Illinois are not as necessary as it tends to follow the national trend when it comes to WNV counts.

Different states have attention on different time scales. Among others, Colorado puts great emphasis on recent disease cases and counts, while California treats each past year roughly equally.

Lastly, the ablation analysis shows the robustness of the models, as well as that each state relies on different sources to predict accurately. This work can be extended by creating a disease operator for the United States. This would entail isolating the temporal variable and isolating an encoding for the position in the United States, and training a transformer architecture to learn the relation between the disease cases, the position, and the time. This would allow predictions at time granularities not supplied as well as locations where data is sparse [17]. The work in this paper will give insight to this operator because the operator will learn the most important features in different parts of the country, and an interpolation between them.

References

[1] Center for disease control, west nile virus. `https://www.cdc.gov/westnile/index.html`.

[2] VectorSurv - vectorborne disease surveillance system. `https://vectorsurv.org`. Accessed: 2023-11-8.

[3] Bowman, C., Gumel, A., Van den Driessche, P., Wu, J., and Zhu, H. A mathematical model for assessing control strategies against west nile virus. *Bulletin of mathematical biology 67*, 5 (2005), 1107–1133.

[4] Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. *Time series analysis: forecasting and control.* John Wiley & Sons, 2015.

[5] de Wit, M. M., Dimas Martins, A., Delecroix, C., Heesterbeek, H., and ten Bosch, Q. Mechanistic models for west nile virus transmission: A systematic review of features, aims, and parameterisation. *bioRxiv* (2023), 2023–10.

[6] DeFelice, N. B., Birger, R., DeFelice, N., Gagner, A., Campbell, S. R., Romano, C., Santoriello, M., Henke, J., Wittie, J., Cole, B., et al. Modeling and surveillance of reporting delays of mosquitoes and humans infected with west nile virus and associations with accuracy of west nile virus forecasts. *JAMA network open 2*, 4 (2019), e193175–e193175.

[7] DeFelice, N. B., Little, E., Campbell, S. R., and Shaman, J. Ensemble forecast of human west nile virus cases and mosquito infection rates. *Nature Communications 8*, 1 (2017), 14592.

[8] DeFelice, N. B., Schneider, Z. D., Little, E., Barker, C., Caillouet, K. A., Campbell, S. R., Damian, D., Irwin, P., Jones, H. M., Townsend, J., et al. Use of temperature to improve west nile virus forecasts. *PLoS computational biology 14*, 3 (2018), e1006047.

[9] Farooq, Z., Rocklöv, J., Wallin, J., Abiri, N., Sewe, M. O., Sjödin, H., and Semenza, J. C. Artificial intelligence to predict west nile virus outbreaks with eco-climatic drivers. *The Lancet Regional Health–Europe 17* (2022).

[10] Gubler, D. J. The global emergence/resurgence of arboviral diseases as public health problems. *Archives of medical research 33*, 4 (2002), 330–342.

[11] Herzen, J., LÃ¤ssig, F., Piazzetta, S. G., Neuer, T., Tafti, L., Raille, G., Pottelbergh, T. V., Pasieka, M., Skrodzki, A., Huguenin, N., Dumonal, M., KoÅcisz, J., Bader, D., Gusset, F., Benheddi, M., Williamson, C., Kosinski, M., Petrik, M., and Grosch, G. Darts: User-friendly modern machine learning for time series. *Journal of Machine Learning Research 23*, 124 (2022), 1–6.

[12] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[13] Holcomb, K. M., Mathis, S., Staples, J. E., Fischer, M., Barker, C. M., Beard, C. B., Nett, R. J., Keyel, A. C., Marcantonio, M., Childs, M. L., et al. Evaluation of an open forecasting challenge to assess skill of west nile virus neuroinvasive disease prediction. *Parasites & Vectors 16*, 1 (2023), 1–13.

[14] Lambert, D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics 34*, 1 (1992), 1–14.

[15] Lanciotti, R., Roehrig, J., Deubel, V., Smith, J., Parker, M., Steele, K., Crise, B., Volpe, K., Crabtree, M., Scherret, J., et al. Origin of the west nile virus responsible for an outbreak of encephalitis in the northeastern united states. *Science 286*, 5448 (1999), 2333–2337.

[16] Lim, B., Arik, S. Ö., Loeff, N., and Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting 37*, 4 (2021), 1748–1764.

[17] Lu, L., Jin, P., and Karniadakis, G. E. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193* (2019).

[18] Manore, C. A., Davis, J. K., Christofferson, R. C., Wesson, D. M., Hyman, J. M., and Mores, C. N. Towards an early warning system for forecasting human west nile virus incidence. *PLoS currents 6* (2014).

[19] Ncei., M. I. Statewide time series. `https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/statewide/time-series`. Accessed: 2023-11-8.

[20] Petersen, L. R., Brault, A. C., and Nasci, R. S. West nile virus: review of the literature. *Jama 310*, 3 (2013), 308–315.

[21] Rodríguez, A., Cui, J., Ramakrishnan, N., Adhikari, B., and Prakash, B. A. Einns: epidemiologically-informed neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2023), vol. 37, pp. 14453–14460.

[22] Shaier, S., Raissi, M., and Seshaiyer, P. Data-driven approaches for predicting spread of infectious diseases through dinns: Disease informed neural networks. *arXiv preprint arXiv:2110.05445* (2021).

[23] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

[24] Wang, L., Chen, J., and Marathe, M. Tdefsi: Theory-guided deep learning-based epidemic forecasting with synthetic information. *ACM Transactions on Spatial Algorithms and Systems (TSAS) 6*, 3 (2020), 1–39.

[25] Watad, A., Watad, S., Mahroum, N., Sharif, K., Amital, H., Bragazzi, N. L., Adawi, M., et al. Forecasting the west nile virus in the united states: an extensive novel data streams–based time series analysis and structural equation modeling of related digital searching behavior. *JMIR public health and surveillance 5*, 1 (2019), e9176.

[26] Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., and Sun, L. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125* (2022).

[27] Wu, H., Xu, J., Wang, J., and Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems 34* (2021), 22419–22430.