



OPEN

## Empirical dynamic modelling and enhanced causal analysis of short-length *Culex* abundance timeseries with vector correlation metrics

Nikos Kollas<sup>1</sup>, Sandra Gewehr<sup>2</sup> & Ioannis Kioutsioukis<sup>1</sup>✉

Employing Empirical Dynamic Modelling we investigate whether model free methods could be applied in the study of *Culex* mosquitoes in Northern Greece. Applying Simplex Projection and S-Map algorithms on yearly timeseries of maximum abundances from 2011 to 2020 we successfully predict the decreasing trend in the maximum number of mosquitoes which was observed in the rural area of Thessaloniki during 2021. Leveraging the use of vector correlation metrics we were able to deduce the main environmental factors driving mosquito abundance such as temperature, rain and wind during 2012 and study the causal interaction between neighbouring populations in the industrial area of Thessaloniki between 2019 and 2020. In all three cases a chaotic and non-linear behaviour of the underlying system was observed. Given the health risk associated with the presence of mosquitoes as vectors of viral diseases these results hint to the usefulness of EDM methods in entomological studies as guides for the construction of more accurate and realistic mechanistic models which are indispensable to public health authorities for the design of targeted control strategies and health prevention measures.

One of the many negative side effects of climate change, among others, is the increase in the number of zoonotic viral diseases transmitted through animal hosts<sup>1</sup>. A notable example of this is the common house mosquito (*Culex* spp.). Several species of the genus *Culex* are known to be principal vectors that can transmit diseases such as the ones caused by the West Nile Virus (WNV)<sup>2</sup> the St. Louis encephalitis virus and the Japanese encephalitis virus for example. In the case of WNV the virus originates in tropical regions and is carried over to other, less warmer, areas by infected birds through migration. By feeding on them, mosquitoes become carriers of the disease which then transmit the pathogen to humans. At all life stages *Culex* spp. are ectothermic and therefore climate sensitive. Studies have suggested that environmental changes, due to an increase in temperature and precipitation caused by global warming, have a significant impact on mosquito populations, facilitating their geographic spread into regions with more temperate climates further north and south<sup>3–9</sup>. A characteristic example is *Culex pipiens*. This is an ubiquitous mosquito species with a close association to humans and a worldwide distribution, inhabiting latitudes as high as Northern Europe and as low as the South Island of New Zealand<sup>10,11</sup>, which is a frequent vector of pathogens of both human and animal diseases<sup>10,12</sup>. In an effort to assist public health authorities in devising appropriate health policies and vector control strategies an increasing number of mathematical models have been developed over the past years aimed at studying mosquito abundances and their dependence upon environmental factors<sup>13–22</sup>.

In this paper we perform a model free analysis of entomological data in the regional unit of Thessaloniki, an area which has seen an increase in the number of WNV reported cases in recent years<sup>23</sup>. The analysis is based on *Empirical Dynamic Modelling* (EDM). This is a data-driven method aimed at reconstructing the attractor manifold of a dynamical system from observations<sup>24–27</sup> capable of assessing the dimensionality and degree of non-linearity of the underlying system. Using EDM it is possible to make short-term predictions of the system's components<sup>28–30</sup> and to infer its causal structure<sup>31–33</sup>. The technique has been applied successfully in various areas of research from paleontology<sup>34–36</sup> and ecology<sup>37–41</sup>, to neuroscience<sup>42–44</sup>, climate studies<sup>45–48</sup> and even

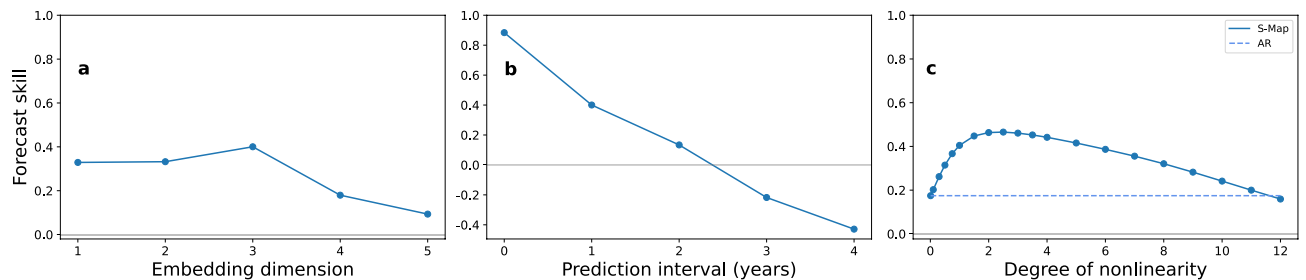
<sup>1</sup>Department of Physics, University of Patras, 26504 Patras, Greece. <sup>2</sup>Ecodevelopment S.A., 57010 Filyro, Greece. ✉email: kioutio@upatras.gr

solar weather prediction<sup>49</sup>. In a previous study, EDM was employed to uncover the environmental variables which drive mosquito abundance of the *Aedes* genus in French Polynesia based on the idea of *Convergent Cross Mapping*<sup>31</sup>. A drawback of this method is that it requires timeseries of a long length which, due to budgetary, time or other constraints, are usually not readily available. To overcome this challenge we apply the method of spatial-State Space Reconstruction (s-SSR)<sup>32,72</sup>, by combining spatial replicates of the variable in question into a composite timeseries, and infer causality from the cross-map skill between variables as a function of time delay between cause and effect. In contradistinction to previous applications we employ *vector correlation* metrics<sup>73</sup> for measuring the quality of cross mapping which provide an enhancement over usual metrics as they are more robust under changes of the embedding dimension<sup>74</sup>.

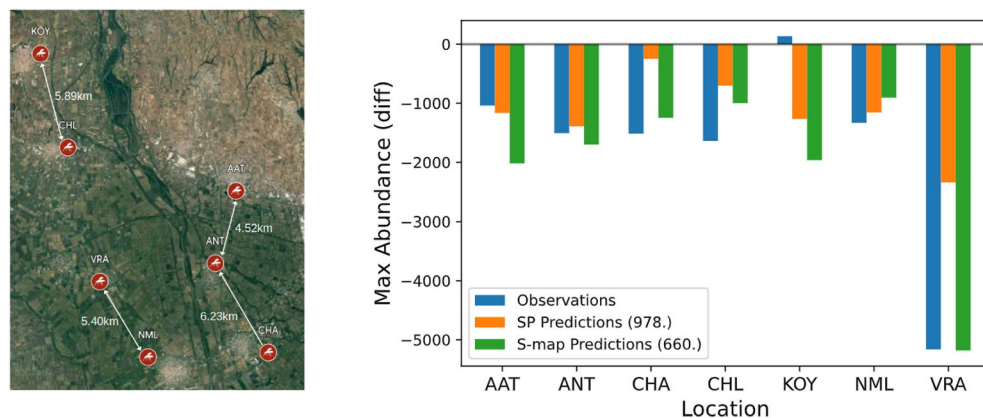
## Results

### Forecasting yearly maximum abundances

As a first test case, we attempt to make out of sample forecasts of the maximum number of mosquitoes expected during the course of a year. The data consists of seven replicates of maximum abundances observed between the months of April and October from 2011 to 2021. Since the mean flight distance of *Culex* mosquitoes encountered in Europe is only of the order of a few hundred meters<sup>50</sup>, the large separation between locations (with a mean nearest neighbour distance of approximately 5.51 kilometers) ensures that the populations are isolated from one another. We therefore apply s-SSR (see Empirical Dynamic Modelling in Methods) to construct a composite state-space. To compensate for annual trends in the data, we first-difference each replicate separately to obtain yearly changes in maximum abundances. To determine the best embedding dimension for making predictions, we run a Simplex Projection (SP) algorithm (see Simplex Projection in Methods) with data from 2012 to 2019 as the input and perform a leave-one-out cross validation method to predict next year's abundance between 2013 and 2020. In Fig. 1a we present the forecast skill of the SP algorithm, as a function of dimension  $E$ . We observe that the forecast skill displays a peak at an embedding dimension  $E = 3$ . For this dimension the forecast skill was higher than 97% of a 1000 randomly generated time-series using Ebisuzaki's method<sup>51</sup> (Table S1 in Supplementary). Projecting further into the future reveals the chaotic behaviour of the system with a decrease in forecast



**Figure 1.** Leave-one-out forecast skill of predicted versus observed yearly differences in maximum mosquito abundance of a SP algorithm (a) as a function of the embedding dimension of the reconstructed state space for predictions made one year into the future and (b) as a function of the prediction interval for an embedding dimension  $E = 3$ . (c) Forecast skill for predictions one year into the future between a local S-Map model and a global AR model as a function of the degree of nonlinearity of the system for an embedding dimension  $E = 3$ .



**Figure 2.** Left: Locations of spatial replicates and closest neighbour distance between sampling stations ANT, AAT, CHA, CHL, KOY, NML and VRA in kilometers (maps made with Google My Maps, Imagery ©2024 TerraMetrics). Right: SP and S-Map predictions of the difference in the maximum number of mosquitoes expected in 2021 compared to 2020 for  $E = 3$  and  $\theta = 2.5$ . The mean absolute error between observations and predictions, is indicated in the labels.

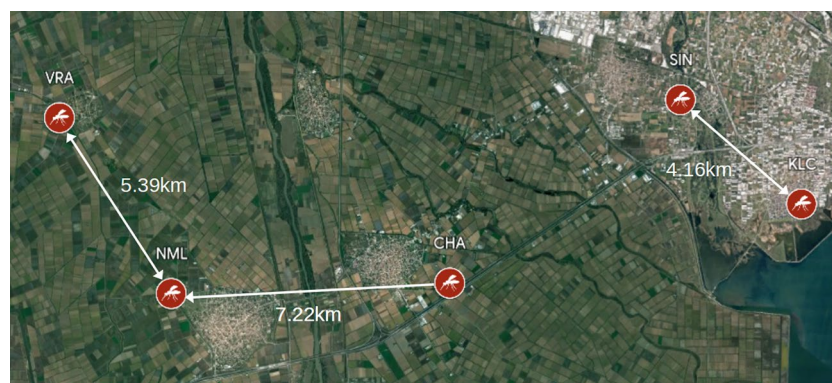
skill and a complete lack in predictability after two years (Fig. 1b). Comparing the forecast skill of a local S-Map model (see S-Map in methods) with that of a global autocorrelation (AR) model of the same dimension as a function of the degree of non-linearity (Fig. 1c) we observe the non-linear nature of the composite time-series with a peak at  $\theta \simeq 2.5$ , larger 97% of the time when compared to a thousand random surrogates.

Using the optimal values of  $E$  and  $\theta$  determined above we now make out-of-sample forecasts of the maximum number of mosquitoes expected in 2021 compared to 2020 for each location. In Fig. 2 we present the predictions of the SP and S-Map algorithms. Both algorithms predict a decrease in the maximum number of mosquitoes expected at each location. Apart from one station (KOY) where a slight increase was actually observed, this was indeed the trend. We observe that the quality of predictions of the S-Map algorithm is better than those given by the SP algorithm with a mean absolute error of 420 and 904 respectively for the stations where prediction and observation match in trend. Excluding KOY from the library and retraining the algorithms with the same parameters shows a slight improvement in predictions with a mean absolute error of 809 for the SP and 403 for the S-Map algorithm, although in this case the optimum degree of non-linearity is found to be equal to  $\theta \simeq 4.5$  (Fig. S1 in Supplementary). Running the algorithm again with the corrected degree increases the mean absolute error to 507 which is worse than that for  $\theta \simeq 2.5$  but still better than the predictions obtained from the SP algorithm (Fig. S2 in Supplementary).

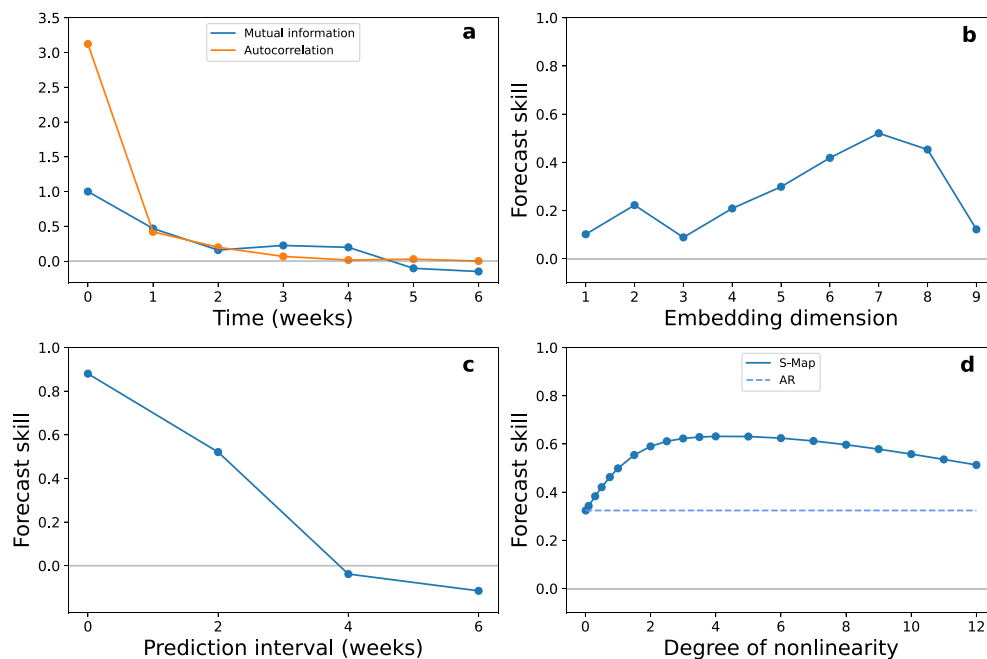
### Causal analysis of environmental effects on mosquito abundance

As a second test case we perform a causal analysis of environmental effects on mosquito populations. From the full set of possible sampling stations (see Data description in Methods), we select a subset of five replicates for which weekly abundances, sampled concurrently from June to September of 2012, were available with 17 consecutive weeks of entries per station (Fig. 3). Once again the large distance between locations (with a mean nearest neighbour distance of 5.59 kilometers) allows us to assume that the stations are independent from each other and we can therefore use s-SSR to construct from them a composite state-space. The environmental variables considered in the causal analysis, include the day mean of the *Land Surface Temperature* (LST), the accumulated rainfall one week before the date of placement (Rain), the mean hourly magnitude of wind (Wind), the *Normalized Difference Vegetation Index* (NDVI), the *Normalized Difference Moisture Index* (NDMI) and the *Normalized Difference Water Index* (NDWI) at each location. In order to increase the density of points in the reconstructed state space and improve the quality of the analysis, min-max normalization for each replicate is performed separately.

Taking into consideration the developmental cycle of *Culex* mosquitoes, which lasts approximately two weeks, we predict with the help of the SP algorithm the state of the system two weeks into the future in order to determine the best embedding dimension. This choice of prediction interval coincides with the first local minimum of the mutual information and is also close to zero for the autocorrelation (Fig. 4a). In panel b of Fig. 4 we present the forecast ability between observed and predicted values of a leave-one-out bootstrap method. We observe that the forecast skill peaks at an embedding dimension  $E = 7$ , which is higher than 99% of a 1000 randomly generated time-series using Ebisuzaki's method (Table S2 in Supplementary). Forecasting further into the future shows a rapid decline in forecast skill (Fig. 4c) with a complete loss of predicting power after four weeks, a behaviour characteristic of a chaotic time-series. The system displays a non-linear behaviour, since the predictions of a local S-Map algorithm are more correlated to their corresponding observed values compared to the predictions of an AR method (Fig. 4d), with a peak at  $\theta \simeq 4$ , higher than 97% of a 1000 surrogate time-series. Although the choice for the best embedding parameters is based on making predictions over a period of two weeks, the reconstructed manifolds contain time-lagged vectors with a lag of only  $\tau = 1$  week. The reason for this choice is the presence of different periods in the dynamics. Mosquito populations are expected to influence each other every fortnight while changes in abundances due to environmental effects can occur on a weekly basis<sup>52</sup>. Performing the above analysis with time-lagged vectors with  $\tau = 2$  weeks instead of one we find that the appropriate embedding dimension in this case has changed and is now equal to  $E = 4$ . Nonetheless the characteristic behaviour of the time series when making predictions further into the future and its non-linearity remain qualitatively the same (Fig. S3



**Figure 3.** Locations of spatial replicates, CHA, KLC, NML, SIN, VRA and closest neighbour distance in kilometers for the causal analysis of environmental effects (maps made with Google My Maps, Imagery ©2024 TerraMetrics).



**Figure 4.** (a) Autocorrelation and mutual information of daily mosquito abundance as a function of time. (b) Leave-one-out forecast skill of SP algorithm as a function of the embedding dimension of the reconstructed state space for predictions two weeks into the future and (c) as a function of the prediction interval for an embedding dimension  $E = 7$ . (d) Leave-one-out forecast skill as a function of the degree of non-linearity of a local S-Map model versus a global AR model for predictions two weeks into the future and an embedding dimension  $E = 7$ .

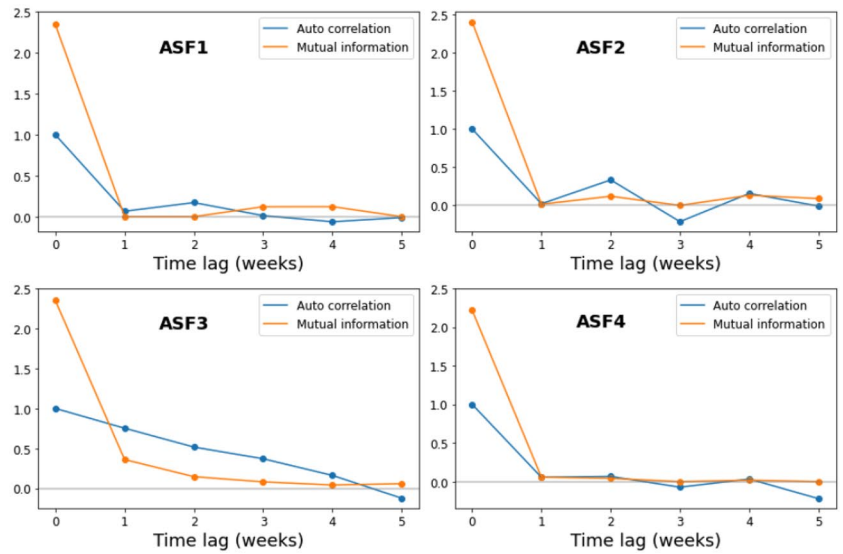
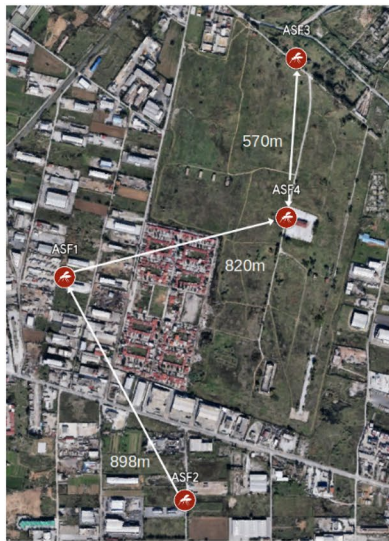
in Supplementary). This is consistent with studies which suggest that what is ultimately important for an EDM analysis is not the value of the embedding dimension  $E$  or of the time-delay  $\tau$  separately but rather the value of the embedding time-window<sup>53–57</sup>  $T_w = (E - 1)\tau$  which in both cases here remains constant and equal to  $T_w = 6$ .

Causality is now inferred by cross-mapping the reconstructed manifolds of each environmental variable with the one obtained from the data on mosquito abundance as a function of the prediction interval (see Cross mapping causal analysis in Methods). In Table 1 we display the average cross-map skill of a vector based correlation metric (See Correlation between random vectors in Methods) for predictions up to one month into the past and the future (for the forecast skill of each mapping as a function of prediction interval, used to calculate the averages, see Fig. S4 in Supplementary). From the positive values we detect LST, Rain and Wind as the main causal factors of daily mosquito abundance, with a mean peak in absolute prediction interval of half a week for temperature and one week for both precipitation and wind (Fig. S5 in Supplementary).

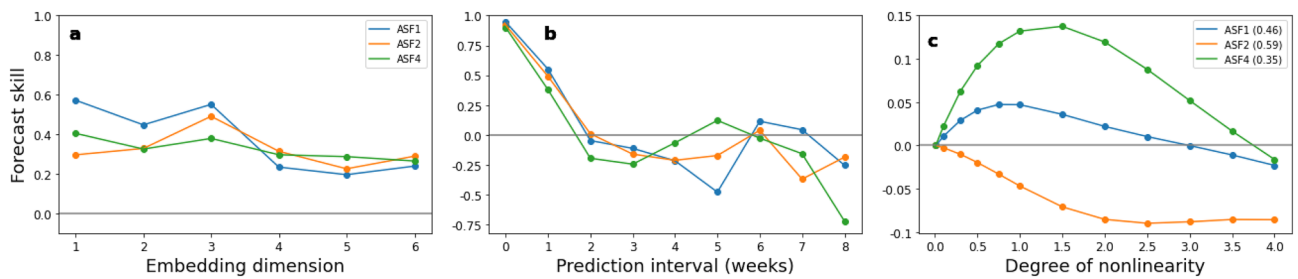
### Causal interaction of neighbouring populations

We now carry out a causal analysis of neighbouring populations. The data in this case consists of four replicates of daily abundances sampled every week between the months of May and September of 2019 and 2020 with 19 and 21 consecutive weeks of entries respectively. In Fig. 5 we display, for each location, the autocorrelation and mutual information of the daily abundance as a function of time. For three out of the four locations (ASF1, ASF2, ASF4) the figure suggests that the choice for the best embedding parameters should be based on predictions made one week into the future. The remaining location (ASF3) shows a gradual decrease in autocorrelation indicating a difference in dynamics so is excluded from the subsequent analysis. To avoid annual trends, we make the data stationary by first-differencing each time-series to obtain weekly changes of abundances. Forecasting each replicate one week into the future we choose  $E = 3$  as the most suitable common embedding dimension (Fig. 6a). The observed forecast skill, for this dimension, compared to the one obtained from a thousand serially-correlated random timeseries generated with Ebisuzaki's method was higher 99%, 88% and 91% of the time (Table S3 in Supplementary). Making predictions further into the future suggests that all three locations exhibit chaotic behaviour, with a decrease in forecast skill and an inability of making any predictions beyond two weeks (Fig. 6b). Running an S-Map analysis of the data as a function of the degree of nonlinearity of the time-series we could detect a non-linear behaviour in only two out of the three locations, ASF1 and ASF4, with  $\theta \simeq 0.75$  (89% of the time higher) and  $\theta \simeq 1.25$  (96% of the time higher) respectively.

In Table 2 we present the average cross-mapping skill of past and future predictions, between locations, for a vector correlation based metric (see Fig. S7 in Supplementary for the forecast skill of each mapping as a function of the prediction interval). We observe that ASF1 causally affects ASF2 with a mean peak in absolute prediction interval of approximately 4 weeks, while at the same time ASF2 affects location ASF4 with a mean peak in absolute prediction interval of 1 week. The latter location is also causally affected by ASF1. The mean absolute peak of 4.5 weeks in the prediction interval suggests in this case the existence of an indirect causal link between the two locations through ASF2.



**Figure 5.** Left: Locations of replicates, ASF1, ASF2, ASF3, ASF4 and closest neighbour distance in meters (maps made with Google My Maps, Imagery ©2024 TerraMetrics). Right: Autocorrelation and mutual information of daily mosquito abundance as a function of time.



**Figure 6.** Leave-one-out forecast skill of SP algorithm (a) as a function of the embedding dimension of the reconstructed state space for predictions one week into the future and (b) as a function of the prediction interval for an embedding dimension  $E = 3$ . (c) Difference in forecast skill for predictions one week into the future between a local S-Map model and a global AR model as a function of the degree of nonlinearity of the system for an embedding dimension  $E = 3$  (the forecast skill of the AR model is indicated in the labels).

	$T_p \leq 0$	$T_p \geq 0$		$T_p \leq 0$	$T_p \geq 0$
Culex xmap LST	0.37 (0 w)	0.31	Culex xmap NDVI	-0.02 (-3 w)	-0.21
LST xmap Culex	0.20	0.37 (1 w)	NDVI xmap Culex	-0.02	0.11 (3 w)
Culex xmap Rain	0.32 (-1 w)	0.26	Culex xmap NDMI	-0.02 (-3 w)	-0.23
Rain xmap Culex	0.45	0.49 (1 w)	NDMI xmap Culex	-0.17	-0.07 (4 w)
Culex xmap Wind	0.27 (-2 w)	0.09	Culex xmap NDWI	-0.13 (-4 w)	-0.20
Wind xmap Culex	0.29	0.33 (0 w)	NDWI xmap Culex	0.18	0.18 (0 w)

**Table 1.** Average cross-map skill of past ( $T_p \leq 0$ ) and future ( $T_p \geq 0$ ) predictions between weekly mosquito abundances and Land Surface Temperature (LST), accumulated rainfall one week before the date of placement (Rain), mean hourly magnitude of wind (Wind), Normalized Difference Vegetation Index (NDVI), Normalized Difference Moisture Index (NDMI) and Normalized Difference Water Index (NDWI) for a vector correlation based metric. The quantities in parentheses indicate the value of the prediction interval (in weeks) where the cross mapping is maximized.

### Discussion

The influence of human activity on the environment, either through increased invasion of animal habitats or changes in climate, is expected to result in an increase in the number of vector borne viral diseases such as yellow fever, Zika, chikungunya and West Nile Virus for example<sup>1</sup>. In light of this threat, techniques for accurately modeling and predicting the abundance of vectors carrying the virus, as in the case of mosquitoes, is important to public health authorities since it will allow them to anticipate possible future outbreaks and take appropriate

	$T_p \leq 0$	$T_p \geq 0$		$T_p \leq 0$	$T_p \geq 0$
ASF1 xmap ASF2	-0.12	0.03 (5 w)	ASF1 xmap ASF4	-0.17	0.00 (3 w)
ASF2 xmap ASF1	0.12 (-3 w)	0.22	ASF4 xmap ASF1	0.08 (-6 w)	-0.11
ASF2 xmap ASF4	-0.19	0.06 (1 w)			
ASF4 xmap ASF2	0.26 (-1 w)	0.14			

**Table 2.** Average cross-map skill of past ( $T_p \leq 0$ ) and future ( $T_p \geq 0$ ) predictions of weekly changes in mosquito abundance between neighboring locations for a vector correlation based metric. The quantities in parentheses indicate the value of the prediction interval (in weeks) where the cross mapping is maximized.

preventive measures. Model free methods make no assumptions about the structure of the system under study and are therefore more general than their mechanistic counterparts. Applying EDM techniques on *Culex* data we were able to correctly predict the decreasing trend in the maximum number of mosquitoes observed during 2021 in a number of stations in the rural area of Thessaloniki. Using the same methodology it was also demonstrated that it is possible to deduce land surface temperature, the accumulated precipitation and the mean hourly wind as the key environmental causal factors driving mosquito abundance. This is consistent with other studies which indicate the same variables as the most important for the dynamics of mosquito populations with time lags similar to the ones reported here<sup>7,14-17,58-65</sup>. Temperature is a particularly important factor as it directly influences the mortality rate and life span of mosquitoes<sup>66-68</sup>. Abundances peak at temperatures between 20 and 30°C, corresponding to a short time in larval developments<sup>2,12</sup>, while high levels of mortality are found outside the 15 to 34°C range. Performing a causal analysis on data in the industrial area of Thessaloniki we also detected interactions between neighboring populations. These were expected due to the short distance between the sampling stations (with a mean distance between closest neighbors of 714.5m) and can often be attributed to non-oriented dispersal of mosquitoes when searching for hosts, food, mates and places for oviposition or shelter<sup>50,69</sup>. Knowledge about interaction dynamics is useful information for the design of intervention strategies. Instead of chemically controlling a larger area for example one could focus attention locally only on those locations which causally affect mosquito abundance in others (such as ASF1 in our example) saving both valuable time and resources.

Though model free methods are advantageous, since they don't suffer from any assumptions that need to be made in the construction of a mechanistic model, one of their main drawbacks is their need for long time-series with 30 or more consecutive observations (as a general rule of thumb). In order to apply these methods in ecological studies, where the available time-series are usually of a short length, it is necessary to combine several of them into a composite state-space using s-SSR. Any inferences based on these models will only therefore apply on a larger scale (depending on the mean closest neighbour distance between replicates) rather than on a local level. Our analysis on the causal effects of environmental factors, for example, suggested the possibility that vegetation, moisture and water content were also causal factors, but these were excluded as such on the basis that the average forecast skill of mapping the abundance manifold onto the manifold obtained from each of these environmental observables was less than zero in both past and future directions. The reason for this can be traced back to the composite nature of the reconstructed state space. Vegetation, moisture and water conditions differ considerably for each location while those for temperature, rainfall and wind are practically the same (Fig. S6 in Supplementary). The inability of making abundance predictions beyond a certain period, apart from the nonlinearity of population dynamics, is also related to the chaotic behaviour of weather. Due to Takens theorem the manifold reconstructed by time-lagged vectors of abundance observations is homeomorphic to the actual state-space of the system which includes environmental variables as degrees of freedom. Any limitations in the forecast ability of the later, with a horizon of approximately two weeks, will similarly affect the forecast ability of abundances. Despite these drawbacks EDM can serve as a useful guide in the construction of more accurate models. In all three cases considered here, each time-series exhibited a chaotic and (with the exception of series ASF2) non-linear behaviour. Comparing the autocorrelation and mutual information of Figs. 4 and 5 we also observe a change of period in the dynamics from two weeks to one week. Any realistic model should be capable to take these effects into account and qualitatively reproduce them. Future research on the relation between the dynamical characteristics (such as embedding dimension, the degree of non-linearity or the optimal time lag) of data used to train mechanistic models with the mathematical structure and choice of parameters of the latter, could possibly assist in that direction.

## Methods

### Data description

Data on mosquito abundances were obtained from EcoDevelopment SA. as part of the Early Warning System for Mosquito borne disease (EYWA) dataset, developed for the EuroGEO Action Group "Earth Observation for Epidemics of Vector-borne Diseases", under the coordination of the National Observatory of Athens/BEYOND Centre of Earth Observation Research and Satellite Remote Sensing. The values of the *Normalized Difference Vegetation Index* (NDVI), the *Normalized Difference Moisture Index* (NDMI), the *Normalized Difference Water Index* (NDWI), the day mean of the *Land Surface Temperature* (LST), the accumulated rainfall one week before the date of placement (Rain) and the mean hourly magnitude of wind (Wind) at each location were obtained from the *Moderate Resolution Imaging Spectroradiometer* (MODIS) instrument aboard the Terra and Aqua satellites.

### Empirical dynamic modelling

EDM is based on Takens's theorem which states that it is possible to construct an embedding of the manifold from lagged vectors of a single observed time-series  $X^{24-27}$

$$\mathbf{X}(t) = \langle X(t), X(t - \tau), \dots, X(t - (E - 1)\tau) \rangle \quad (1)$$

where  $\tau$  is a *time-lag* which can be determined by the first zero of the autocorrelation function or the first local minimum of the mutual information<sup>70,71</sup>. Taking advantage of the topological structure of the reconstructed 'shadow' manifold it is possible to make predictions about the evolution of the system<sup>28-30</sup> and infer its causal structure<sup>31-33</sup>. For series of a short length, as is usually the case in ecological studies, one can construct a composite state-space from spatial replicates using *spatial State Space Reconstruction*<sup>32,72</sup> (s-SSR), discarding from the analysis any embedded vectors with overlap between their components.

### Simplex projection

Simplex projection<sup>28</sup> is a *k-nearest neighbours* regression algorithm. For any observation  $\mathbf{X}(t)$  one can predict a feature  $F(\mathbf{X}(t))$  as the weighted average of its  $E + 1$  closest neighbours

$$\hat{F}(\mathbf{X}(t)) = \sum_{i=1}^{E+1} w_i F(\mathbf{X}(t_i)), \quad (2)$$

where

$$w_i = \frac{\exp[-d_i/\bar{d}]}{\sum_j \exp[-d_j/\bar{d}]}, \quad i = 1, \dots, E + 1 \quad (3)$$

is the weight of the  $i$ -th closest neighbour  $\mathbf{X}(t_i)$  to  $\mathbf{X}(t)$  with euclidean distance  $d_i$ , and  $\bar{d}$  is the mean distance from all of its neighbours. Choosing  $F(\mathbf{X}(t_i)) = X(t_i + T_p)$  in Eq. (2) one can employ the algorithm to make forecasts, by tracking where each neighbour will end up after  $T_p$  time steps in the reconstructed state space. Making predictions one time step into the future it is possible to get an estimate for the best *embedding dimension*  $E$  by requiring the cross-validation between observed and predicted time-series to be maximized<sup>28</sup>. The same algorithm can also be used to distinguish between measurement error and chaos in the time-series. Making predictions further into the future for a chaotic system results in a sharp decrease in forecast skill compared to a noisy series for which the ability of making predictions as a function of the prediction interval  $T_p$  remains relatively constant<sup>28-30</sup>.

### S-map

The S-Map algorithm can be used to test for non-linearity in a candidate time-series<sup>30</sup>. For every observation  $\mathbf{X}(t)$ , an  $E + 1$  order autoregressive model is employed to forecast the value of the time-series  $T_p$  time steps ahead as the scalar product of an  $E + 1$  dimensional vector of coefficients  $\mathbf{C}_t$  with  $\mathbf{X}(t)$

$$\hat{X}(t + T_p) = \mathbf{C}_t \cdot \mathbf{X}(t). \quad (4)$$

The model is trained on a locally weighted set of lagged vectors

$$\left\{ \tilde{\mathbf{X}}(t') = e^{-\theta d(t')/\bar{d}} \mathbf{X}(t'), \quad t \neq t' \right\}, \quad \theta \geq 0 \quad (5)$$

where  $d(t')$  is the euclidean distance of  $\mathbf{X}(t)$  from  $\mathbf{X}(t')$  and  $\bar{d}$  is the mean distance. If  $\theta > 0$  then the training set is different for each prediction, with vectors closer to  $\mathbf{X}(t)$  contributing more to the model. This corresponds to a different vector of coefficients  $\mathbf{C}_t$  each time which can be calculated through a singular value decomposition or a least-squares approach. For  $\theta = 0$ ,  $\mathbf{C}_t$  is the same for every prediction and the model is global and equal to an autoregression algorithm of the same order. If the quality of predictions improves with the local model compared to the predictions obtained from the global model then the time-series can be considered to be non-linear with the degree of non-linearity represented by the parameter  $\theta$ .

### Cross mapping causal analysis

If  $X$  and  $Y$  are part of the same dynamical system then it is possible to infer whether a cause and effect relationship exists between the two variables by cross mapping their respective 'shadow' manifolds constructed from time lagged vectors of their observations<sup>31-33</sup>. By choosing  $F(\mathbf{X}(t_i)) = Y(t_i + T_p)$  as a feature of the Simplex algorithm, it is possible to predict  $Y$  from  $X$  (in this case the prediction interval  $T_p$  can take both positive as well as negative values). Causality can now be ascertained from the correlation between predicted and observed values as a function of  $T_p$ . If the mean correlation for positive values of  $T_p$  is greater than that for negative values this means that past values of  $X$  are better at predicting future values of  $Y$  so it can be inferred that the former variable causally affects the latter,  $X \prec Y$ . If on the other hand the mean correlation for negative values of  $T_p$  is greater than that for positive values then  $Y \prec X$ .

### Correlation between random vectors

Setting  $F(\mathbf{X}(t_i)) = \mathbf{Y}(t_i + T_p)$  as a feature of the Simplex algorithm, it is possible to predict the full lagged vector of  $Y$ , not only its first component. In order to assess the quality of predictions  $\hat{\mathbf{Y}}$  in this case we employ the following linear correlation coefficient which is suitable for use with random vectors

$$\rho(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{\text{tr}(\Sigma_{\mathbf{Y}\hat{\mathbf{Y}}})}{\text{tr}(\sqrt{\Sigma_{\mathbf{Y}\mathbf{Y}}\Sigma_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}})} \quad (6)$$

where  $\Sigma_{\mathbf{X}\mathbf{Y}}$  denotes the covariance matrix between random vectors  $\mathbf{X}$  and  $\mathbf{Y}$ . It is proven that the above definition, which is a measure of the mean euclidean distance between the two sets of vectors, satisfies all of the properties of Pearson's correlation coefficient<sup>73</sup> which it reduces to in the univariate case. Recently, it was shown that the use of vector metrics provides a marked improvement of the causal structure of a non-linear system over other metrics<sup>74</sup>.

### Data availability

The data that support the findings of this study are available from Ecodevelopment S.A. but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the corresponding author upon reasonable request and with permission of Ecodevelopment S.A.

Received: 22 September 2023; Accepted: 8 February 2024

Published online: 13 February 2024

### References

1. Failloux, A.-B. Human activities and climate change in the emergence of vector-borne diseases. *Comptes Rendus Biologies* **342**, 269–270. <https://doi.org/10.1016/j.crv.2019.09.023> (2019). Insects: Friends, foes, and models / Insectes : amis, ennemis et modèles.
2. Moser, S. K. *et al.* Scoping review of *Culex* mosquito life history trait heterogeneity in response to temperature. *Parasites Vectors* **16**, 200. <https://doi.org/10.1186/s13071-023-05792-3> (2023).
3. Hongoh, V., Berrang-Ford, L., Scott, M. & Lindsay, L. Expanding geographical distribution of the mosquito, *Culex pipiens*, in Canada under climate change. *Appl. Geogr.* **33**, 53–62. <https://doi.org/10.1016/j.apgeog.2011.05.015> (2012).
4. Morin, C. W. & Comrie, A. C. Regional and seasonal response of a West Nile virus vector to climate change. *Proc. Natl. Acad. Sci.* **110**, 15620–15625. <https://doi.org/10.1073/pnas.1307135110> (2013).
5. Paz, S. Climate change impacts on West Nile virus transmission in a global context. *Philos. Trans. R. Soc. B: Biol. Sci.* **370**, 20130561. <https://doi.org/10.1098/rstb.2013.0561> (2015).
6. Samy, A. M. *et al.* Climate change influences on the global potential distribution of the mosquito *Culex quinquefasciatus*, vector of west Nile virus and lymphatic filariasis. *PLoS ONE* **11**, e0163863 (2016).
7. Ruybal, J. E., Kramer, L. D. & Kilpatrick, A. M. Geographic variation in the response of *Culex pipiens* life history traits to temperature. *Parasites Vectors* **9**, 116. <https://doi.org/10.1186/s13071-016-1402-z> (2016).
8. Watts, M. J. & Sarto i Monteys, V., Mortyn, P. G. & Kotsila, P., The rise of West Nile virus in southern and southeastern Europe: A spatial-temporal analysis investigating the combined effects of climate, land use and economic changes. *One Health* **13**, 100315. <https://doi.org/10.1016/j.onehlt.2021.100315> (2021).
9. Ewing, D. A., Purse, B. V., Cobbold, C. A. & White, S. M. A novel approach for predicting risk of vector-borne disease establishment in marginal temperate environments under climate change: West Nile virus in the UK. *J. R. Soc. Interface* **18**, 20210049. <https://doi.org/10.1098/rsif.2021.0049> (2021).
10. Farajollahi, A., Fonseca, D. M., Kramer, L. D. & Marm Kilpatrick, A. Bird biting mosquitoes and human disease: A review of the role of *Culex pipiens* complex mosquitoes in epidemiology. *Infect. Genet. Evol.* **11**, 1577–1585. <https://doi.org/10.1016/j.meegid.2011.08.013> (2011).
11. Booth, M. Chapter three - climate change and the neglected tropical diseases. vol. 100 of *Advances in Parasitology*, 39–126. <https://doi.org/10.1016/bs.apar.2018.02.001> (Academic Press, 2018).
12. Ciota, A. T., Matarachio, A. C., Kilpatrick, A. M. & Kramer, L. D. The effect of temperature on life history traits of *Culex* mosquitoes. *J. Med. Entomol.* **51**, 55–62. <https://doi.org/10.1603/ME13003> (2014).
13. Danforth, M. E., Reisen, W. K. & Barker, C. M. The impact of cycling temperature on the transmission of West Nile virus. *J. Med. Entomol.* **53**, 681–686. <https://doi.org/10.1093/jme/tjw013> (2016).
14. Stilianakis, N. I. *et al.* Identification of climatic factors affecting the epidemiology of human West Nile virus infections in northern Greece. *PLoS ONE* **11**, 1–17. <https://doi.org/10.1371/journal.pone.0161510> (2016).
15. Moirano, G. *et al.* West Nile virus infection in northern Italy: Case-crossover study on the short-term effect of climatic parameters. *Environ. Res.* **167**, 544–549. <https://doi.org/10.1016/j.envres.2018.08.016> (2018).
16. Marini, G. *et al.* West Nile virus transmission and human infection risk in Veneto (Italy): A modelling analysis. *Sci. Rep.* **8**, 14005. <https://doi.org/10.1038/s41598-018-32401-6> (2018).
17. Kioutsoukis, I. & Stilianakis, N. I. Assessment of West Nile virus transmission risk from a weather-dependent epidemiological model and a global sensitivity analysis framework. *Acta Trop.* **193**, 129–141. <https://doi.org/10.1016/j.actatropica.2019.03.003> (2019).
18. Calzolari, M. *et al.* Enhanced West Nile virus circulation in the Emilia-Romagna and Lombardy regions (Northern Italy) in 2018 detected by entomological surveillance. *Front. Vet. Sci.* <https://doi.org/10.3389/fvets.2020.00243> (2020).
19. Angelou, A., Kioutsoukis, I. & Stilianakis, N. I. A climate-dependent spatial epidemiological model for the transmission risk of West Nile virus at local scale. *One Health* **13**, 100330. <https://doi.org/10.1016/j.onehlt.2021.100330> (2021).
20. Fasano, A. *et al.* An epidemiological model for mosquito host selection and temperature-dependent transmission of West Nile virus. *Sci. Rep.* **12**, 19946. <https://doi.org/10.1038/s41598-022-24527-5> (2022).
21. Tsantalidou, A. *et al.* Mammoth: An earth observational data-driven model for mosquitoes abundance prediction. *Remote Sensing* <https://doi.org/10.3390/rs13132557> (2021).
22. Ferraccioli, F. *et al.* Effects of climatic and environmental factors on mosquito population inferred from West Nile virus surveillance in Greece. *Sci. Rep.* **13**, 18803. <https://doi.org/10.1038/s41598-023-45666-3> (2023).
23. Tsioka, K. *et al.* West Nile virus in *Culex* mosquitoes in Central Macedonia, Greece, 2022. *Viruses* <https://doi.org/10.3390/v15010224> (2023).
24. Packard, N. H., Crutchfield, J. P., Farmer, J. D. & Shaw, R. S. Geometry from a time series. *Phys. Rev. Lett.* **45**, 712–716. <https://doi.org/10.1103/PhysRevLett.45.712> (1980).



25. Takens, F. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980* (eds Rand, D. & Young, L.-S.) 366–381 (Springer, 1981).
26. Sauer, T., Yorke, J. A. & Casdagli, M. Embedology. *J. Stat. Phys.* **65**, 579–616. <https://doi.org/10.1007/BF01053745> (1991).
27. Deyle, E. R. & Sugihara, G. Generalized theorems for nonlinear state space reconstruction. *PLoS ONE* **6**, 1–8. <https://doi.org/10.1371/journal.pone.0018295> (2011).
28. Sugihara, G. & May, R. M. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature* **344**, 734–741. <https://doi.org/10.1038/344734a0> (1990).
29. Sugihara, G. *et al.* Distinguishing error from chaos in ecological time series. *Philos. Trans. R. Soc. London Series B Biol. Sci.* **330**, 235–251. <https://doi.org/10.1098/rstb.1990.0195> (1990).
30. Sugihara, G., Grenfell, B. T., May, R. M. & Tong, H. Nonlinear forecasting for the classification of natural time series. *Philos. Trans. R. Soc. London Series A Phys. Eng. Sci.* **348**, 477–495. <https://doi.org/10.1098/rsta.1994.0106> (1994).
31. Sugihara, G. *et al.* Detecting causality in complex ecosystems. *Science* **338**, 496–500. <https://doi.org/10.1126/science.1227079> (2012).
32. Clark, A. T. *et al.* Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology* **96**, 1174–1181. <https://doi.org/10.1890/14-1479.1> (2015).
33. Ye, H., Deyle, E. R., Gilarranz, L. J. & Sugihara, G. Distinguishing time-delayed causal interactions using convergent cross mapping. *Sci. Rep.* **5**, 14750. <https://doi.org/10.1038/srep14750> (2015).
34. Hannisdal, B., Haaga, K. A., Reitan, T., Diego, D. & Liow, L. H. Common species link global ecosystems to climate change: dynamical evidence in the planktonic fossil record. *Proc. R. Soc. B: Biol. Sci.* **284**, 20170722. <https://doi.org/10.1098/rspb.2017.0722> (2017).
35. Cermeño, P., Benton, M. J., Paz, O. & Vèrard, C. Trophic and tectonic limits to the global increase of marine invertebrate diversity. *Sci. Rep.* **7**, 15969. <https://doi.org/10.1038/s41598-017-16257-w> (2017).
36. Cramer, K. L., O’Dea, A., Carpenter, C. & Norris, R. D. A 3000 year record of Caribbean reef urchin communities reveals causes and consequences of long-term decline in *Diadema antillarum*. *Ecography* **41**, 164–173. <https://doi.org/10.1111/ecog.02513> (2018).
37. Deyle, E. R. *et al.* Predicting climate effects on pacific sardine. *Proc. Natl. Acad. Sci.* **110**, 6430–6435. <https://doi.org/10.1073/pnas.1215506110> (2013).
38. Ye, H. *et al.* Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proc. Natl. Acad. Sci.* **112**, E1569–E1576. <https://doi.org/10.1073/pnas.1417063112> (2015).
39. Deyle, E. R., May, R. M., Munch, S. B. & Sugihara, G. Tracking and forecasting ecosystem interactions in real time. *Proc. R. Soc. B: Biol. Sci.* **283**, 20152258. <https://doi.org/10.1098/rspb.2015.2258> (2016).
40. Ushio, M. *et al.* Fluctuating interaction network and time-varying stability of a natural fish community. *Nature* **554**, 360–363. <https://doi.org/10.1038/nature25504> (2018).
41. Rogers, T. L. *et al.* Trophic control changes with season and nutrient loading in lakes. *Ecol. Lett.* **23**, 1287–1297. <https://doi.org/10.1111/ele.13532> (2020).
42. McBride, J. C. *et al.* Sugihara causality analysis of scalp EEG for detection of early Alzheimer’s disease. *NeuroImage Clin.* **7**, 258–265. <https://doi.org/10.1016/j.nicl.2014.12.005> (2015).
43. Tajima, S., Yanagawa, T., Fujii, N. & Toyoizumi, T. Untangling brain-wide dynamics in consciousness by cross-embedding. *PLoS Comput. Biol.* **11**, 1–28. <https://doi.org/10.1371/journal.pcbi.1004537> (2015).
44. Watanakeesuntorn, W. *et al.* Massively parallel causal inference of whole brain dynamics at single neuron resolution. In *2020 IEEE 26th International Conference on Parallel and Distributed Systems (ICPADS)*, 196–205, <https://doi.org/10.1109/ICPADS51040.2020.00035> (2020).
45. Tsonis, A. A. *et al.* Dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature. *Proc. Natl. Acad. Sci.* **112**, 3253–3256. <https://doi.org/10.1073/pnas.1420291112> (2015).
46. van Nes, E. H. *et al.* Causal feedbacks in climate change. *Nat. Clim. Chang.* **5**, 445–448. <https://doi.org/10.1038/nclimate2568> (2015).
47. Stathopoulos, S., Tsonis, A. A. & Kourtidis, K. On the cause-and-effect relations between aerosols, water vapor, and clouds over East Asia. *Theoret. Appl. Climatol.* **144**, 711–722. <https://doi.org/10.1007/s00704-021-03563-7> (2021).
48. Díaz, E., Adsuara, J. E., Martínez, Á. M., Piles, M. & Camps-Valls, G. Inferring causal relations from observational long-term carbon and water fluxes records. *Sci. Rep.* **12**, 1610. <https://doi.org/10.1038/s41598-022-05377-7> (2022).
49. Sarp, V., Kilcik, A., Yurchyshyn, V., Rozelot, J. P. & Ozguc, A. Prediction of solar cycle 25: A non-linear approach. *Mon. Not. R. Astron. Soc.* **481**, 2981–2985. <https://doi.org/10.1093/mnras/sty2470> (2018).
50. Verdonschot, P. F. & Besse-Lototskaya, A. A. Flight distance of mosquitoes (culicidae): A metadata analysis to support the management of barrier zones around rewetted and newly constructed wetlands. *Limnologia* **45**, 69–79. <https://doi.org/10.1016/j.limno.2013.11.002> (2014).
51. Ebisuzaki, W. A method to estimate the statistical significance of a correlation when the data are serially correlated. *J. Clim.* **10**, 2147–2153. [https://doi.org/10.1175/1520-0442\(1997\)010<2147:AMTETS>2.0.CO;2](https://doi.org/10.1175/1520-0442(1997)010<2147:AMTETS>2.0.CO;2) (1997).
52. Crespo-Miguel, R. & Cao-García, F. J. Predictability of population fluctuations. *Mathematics* <https://doi.org/10.3390/math10173176> (2022).
53. Broomhead, D. & King, G. P. Extracting qualitative dynamics from experimental data. *Physica D* **20**, 217–236. [https://doi.org/10.1016/0167-2789\(86\)90031-X](https://doi.org/10.1016/0167-2789(86)90031-X) (1986).
54. Gibson, J. F., Doynne Farmer, J., Casdagli, M. & Eubank, S. An analytic approach to practical state space reconstruction. *Physica D Nonlinear Phenom.* **57**, 1–30. [https://doi.org/10.1016/0167-2789\(92\)90085-2](https://doi.org/10.1016/0167-2789(92)90085-2) (1992).
55. Rosenstein, M. T., Collins, J. J. & De Luca, C. J. Reconstruction expansion as a geometry-based framework for choosing proper delay times. *Physica D* **73**, 82–98. [https://doi.org/10.1016/0167-2789\(94\)90226-7](https://doi.org/10.1016/0167-2789(94)90226-7) (1994).
56. Kugiumtzis, D. State space reconstruction parameters in the analysis of chaotic time series: The role of the time window length. *Physica D* **95**, 13–28. [https://doi.org/10.1016/0167-2789\(96\)00054-1](https://doi.org/10.1016/0167-2789(96)00054-1) (1996).
57. Small, M. & Tse, C. Optimal embedding parameters: A modelling paradigm. *Physica D* **194**, 283–296. <https://doi.org/10.1016/j.physd.2004.03.006> (2004).
58. Paz, S. & Albersheim, I. Influence of warming tendency on *Culex pipiens* population abundance and on the probability of West Nile fever outbreaks (Israeli case study: 2001–2005). *EcoHealth* **5**, 40–48. <https://doi.org/10.1007/s10393-007-0150-0> (2008).
59. Bisanzio, D. *et al.* Spatio-temporal patterns of distribution of West Nile virus vectors in eastern piedmont region, Italy. *Parasites Vectors* **4**, 230. <https://doi.org/10.1186/1756-3305-4-230> (2011).
60. Wang, J., Ogden, N. H. & Zhu, H. The impact of weather conditions on *Culex pipiens* and *Culex restuans* (Diptera: Culicidae) abundance: A case study in peel region. *J. Med. Entomol.* **48**, 468–475. <https://doi.org/10.1603/ME10117> (2011).
61. Lebl, K., Brugger, K. & Rubel, F. Predicting *Culex pipiens/restuans* population dynamics by interval lagged weather data. *Parasites Vectors* **6**, 129. <https://doi.org/10.1186/1756-3305-6-129> (2013).
62. Bravo-Barriga, D. *et al.* The mosquito fauna of the western region of Spain with emphasis on ecological factors and the characterization of *Culex pipiens* forms. *J. Vector Ecol.* **42**, 136–147. <https://doi.org/10.1111/jvec.12248> (2017).
63. Soh, S. & Aik, J. The abundance of *Culex* mosquito vectors for West Nile virus and other flaviviruses: A time-series analysis of rainfall and temperature dependence in singapore. *Sci. Total Environ.* **754**, 142420. <https://doi.org/10.1016/j.scitotenv.2020.142420> (2021).

64. Koenraadt, C. & Harrington, L. C. Flushing effect of rain on container-inhabiting mosquitoes *Aedes aegypti* and *Culex pipiens* (Diptera: Culicidae). *J. Med. Entomol.* **45**, 28–35. <https://doi.org/10.1093/jmedent/45.1.28> (2008).
65. Jones, C. E., Lounibos, L. P., Marra, P. P. & Kilpatrick, A. M. Rainfall influences survival of *Culex pipiens* (Diptera: Culicidae) in a residential neighborhood in the mid-atlantic United States. *J. Med. Entomol.* **49**, 467–473. <https://doi.org/10.1603/ME11191> (2012).
66. Su, T. & Mulla, M. Effects of temperature on development, mortality, mating and blood feeding behavior of *Culiseta incidens* (diptera: Culicidae). *J. Vector Ecol. J. Soc. Vector Ecol.* **26**, 83–92 (2001).
67. Debat, V., Béugin, M., Legout, H. & David, J. R. allometric and nonallometric components of *Drosophila* wing shape respond differently to developmental temperature. *Evolution* **57**, 2773–2784. <https://doi.org/10.1111/j.0014-3820.2003.tb01519.x> (2003).
68. Gunay, F., Alten, B. & Ozsoy, E. D. Narrow-sense heritability of body size and its response to different developmental temperatures in *Culex quinquefasciatus* (say 1923). *J. Vector Ecol.* **36**, 348–354. <https://doi.org/10.1111/j.1948-7134.2011.00175.x> (2011).
69. Service, M. W. Mosquito (Diptera: Culicidae) dispersal: The long and short of it. *J. Med. Entomol.* **34**, 579–588. <https://doi.org/10.1093/jmedent/34.6.579> (1997).
70. Fraser, A. M. & Swinney, H. L. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A* **33**, 1134–1140. <https://doi.org/10.1103/PhysRevA.33.1134> (1986).
71. Krakovská, A., Mezeiová, K. & Budáčová, H. Use of false nearest neighbours for selecting variables and embedding parameters for state space reconstruction. *J. Complex Syst.* **2015**, 932750. <https://doi.org/10.1155/2015/932750> (2015).
72. Hsieh, C., Anderson, C. & Sugihara, G. Extending nonlinear analysis to short ecological time series. *Am. Nat.* **171**, 71–80. <https://doi.org/10.1086/524202> (2008) (PMID: 18171152).
73. Puccetti, G. Measuring linear correlation between random vectors. *Inf. Sci.* **607**, 1328–1347. <https://doi.org/10.1016/j.ins.2022.06.016> (2022).
74. Kollas, N., Gewehr, S., Mourelatos, S. & Kioutsoukis, I. An improved indicator for causal interaction in non-linear systems. *Environ. Sci. Proc.* <https://doi.org/10.3390/envirosciproc2023026092> (2023).

## Acknowledgements

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH - CREATE - INNOVATE (project code: T2EAK-02070). This work was also supported from the EIC Horizon Prize “Early Warning for Epidemics”.

## Author contributions

Conceptualization, N.K. and I.K.; methodology, N.K.; data curation, S.G.; writing-original draft preparation, N.K.; writing-review and editing, N.K. and I.K.; supervision, I.K.; funding acquisition, I.K. All authors have read and reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-54054-4>.

**Correspondence** and requests for materials should be addressed to I.K.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024, corrected publication 2024