Causality analyses

Convergent cross mapping

The convergent cross mapping (CCM) method (Sugihara et al. 2012; BozorgMagham et al. 2015) was used to investigate the causal relationships within this biological system and was based on transferring information from the driver time-series to the response time-series under the assumption of directional causality (from driver to response). In this causal scenario, the response time-series necessarily contains signatures (information) about the driver time-series whereas the reverse may not be true. The CCM method uses time-lagged components of the response time-series to estimate the dynamics of the candidate driver time-series. A better estimate of the driver behavior shows a stronger causal influence on the response variable. In addition, if the two variables are dynamically connected, a better estimate of the driver signal would be expected from a larger number of observations, referred to as the library of the time-series. To obtain a quantitative measure of causality, the Pearson correlation coefficient between the estimated and the original driver signals was used. In addition, to avoid spurious localized (short-term) correlated dynamics between a candidate driver and response, the recovery of the driver signal was investigated as a function of library length L. The library length describes the number of historical observations that are used to generate estimations, and can be a subset of the total number of observations N.

The first step in implementing the CCM method for two time-series $x(t)$ and $y(t)$ (the driver and response signal, respectively) was to generate the reconstructed phase spaces (shadowing manifolds) from the libraries of the two time-series with length L data points:

$$\{X^i{}_L\} = \{x(i), x(i+1), \ldots, x(i+L-1)\} \quad (1)$$

$$\{Y^i{}_L\} = \{y(i), y(i+1), \ldots, y(i+L-1)\} \quad (2)$$

for $i = 1$ to $i = N + 1 - L$ where N is the number of data points in the time-series and the superscript "i" denotes the i-th library. The libraries must sweep the entire length of the original time series (see **Fig. S6**). A reconstructed phase space was generated by using a proper time lag (τ) and an embedding dimension (E). The average mutual information measure was used to select the time lag (Abarbanel 1996). The embedding dimension is a measure of the number of observations used for estimation. The false neighborhood method was used to determine an optimal value for E (Cao 1997). The time-delayed vectors of the reconstructed phase spaces were:

$$X_k = (x(k), x(k-\tau), ..., x(k-(E-1)\tau)) \quad (3)$$

$$Y_k = (y(k), y(k-\tau), ..., y(k-(E-1)\tau)) \quad (4)$$

for $k = 1 + (E-1)\tau$ to $k = L$ where the subscript k shows the k-th point of the i-th library of length L data points. Based on the spore release and meteorological signals, a common time lag of τ = 2 hours and embedding dimensions E = 6 was found.

After the reconstruction of the phase spaces from the selected libraries, for each E-dimensional point in the response reconstructed phase space (a generic E-dimensional point was denoted as Y-central in **Fig. S7**), a sufficient number of nearest neighbor points were selected and their distances, d_i, to the Y-central point were determined. Next, the contemporaneous of each neighbor point in the driver reconstructed phase space was determined. The spatial average of the designated points in the driver shadowing manifold (shown by a star in **Fig. S7**) was determined by using d_i's as the weighting factors (Sugihara et al. 2012). This procedure was repeated for all the E-dimensional points in the response reconstructed phase space, and the correlation between the resultant points and the X-central points, the contemporaneous of the Y-central points in the X_k, was measured.

The CCM coefficient, ρ, as a function of library length L of hourly observations was defined as the average of the Pearson correlation coefficients corresponding to the libraries with the specified

length. Causality is indicated by a CCM coefficient ρ that increases significantly with increasing library length and is significantly greater than zero for large library length. Higher values of the CCM coefficients indicate stronger causal influence. In this study, the convergence and relative magnitude of ρ was investigated using spore concentration as the response signal and meteorological variables as the candidate driver signals.

Multivariate state space forecasting

The multivariate state space forecasting method (Deyle et al. 2013) was inspired by a conceptual combination of the Granger causality method (Granger 1969) and the simplex method for predicting the short term evolution of deterministic chaotic time-series (Farmer and Sidorowich 1987). The multivariate forecasting method augments the information of a driver with the information of the response signal and exploits the cumulative information for a better prediction. This study applied the multivariate forecasting method and expected to observe significant improvement in the prediction of the spore concentration when the information of an influential meteorological variable, which was detected by the CCM method, was augmented with the information of the spore concentration.

In this analysis, the reconstructed phase space of the spore concentration was the same as Y_k, introduced in equation (4). The vectors of Y_k were augmented with the data of an environmental signal represented by $x(t)$. The augmented time-delayed vector was:

$$Y_{augmented} = (y(k), y(k-\tau), ..., y(k-(E-2)\tau, x(k-\Delta))) \quad (5)$$

where Δ shows the delay between the actuation of the driver and the response of the system.

We used Y_k and $Y_{augmented}$ for short term (maximum 12 hours lead time) single variable and multivariate forecasting of the spore concentration, respectively (Deyle et al. 2013; Farmer and Sidorowich 1987). The statistical significance of the improvement of the root mean square (RMS) of the forecasting error between the multivariate and single variable forecasting schemes was investigated to

verify the effectiveness of augmentation of the environmental data. The forecasting error in each case, multivariate and single variable, was defined as the difference between the observed spore concentration (correct values) and the forecast results.

The hourly observed spore concentrations were denoted by Y_{obs}, the hourly forecasts by Y_f and the number of available hourly forecasts by n. The RMS of the forecasting error, S, was defined as:

$$S = \sqrt{\frac{1}{n}\sum (Y_{obs} - Y_f)^2} \quad (6)$$

The number of possible hourly forecasts depends on the starting point, the lead time and the length of the time series. The starting point selected was 0700 on 11 May 2012. This provided *n*= 65 in the RMS calculations and sufficient record for forecasting purposes.

References

Abarbanel, H. (1996). *Analysis of observed chaotic data*. Berlin: Springer-Verlag.

BozorgMagham, A. E., Motesharrei, S., Penny, S. G., & Kalnay, E. (2015). Causality analysis: Identifying the leading element in a coupled dynamical system. *PLos One, 10*(6), e0131226, doi:10.1371/journal.pone.0131226.

Cao, L. (1997). Practical method for determining the minimum embedding dimension of a scalar time series. *Physica D: Nonlinear Phenomena, 110*(1), 43-50.

Deyle, E. R., Fogarty, M., Hsieh, C.-h., Kaufman, L., MacCall, A. D., Munch, S. B., et al. (2013). Predicting climate effects on Pacific sardine. *Proceedings of the National Academy of Sciences, 110*(16), 6430-6435.

Farmer, J. D., & Sidorowich, J. J. (1987). Predicting chaotic time series. *Physical Review Letters, 59*(8), 845.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424-438.

Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., et al. (2012). Detecting causality in complex ecosystems. *Science, 338*(6106), 496-500, doi:10.1126/science.1227079.

Supporting Information

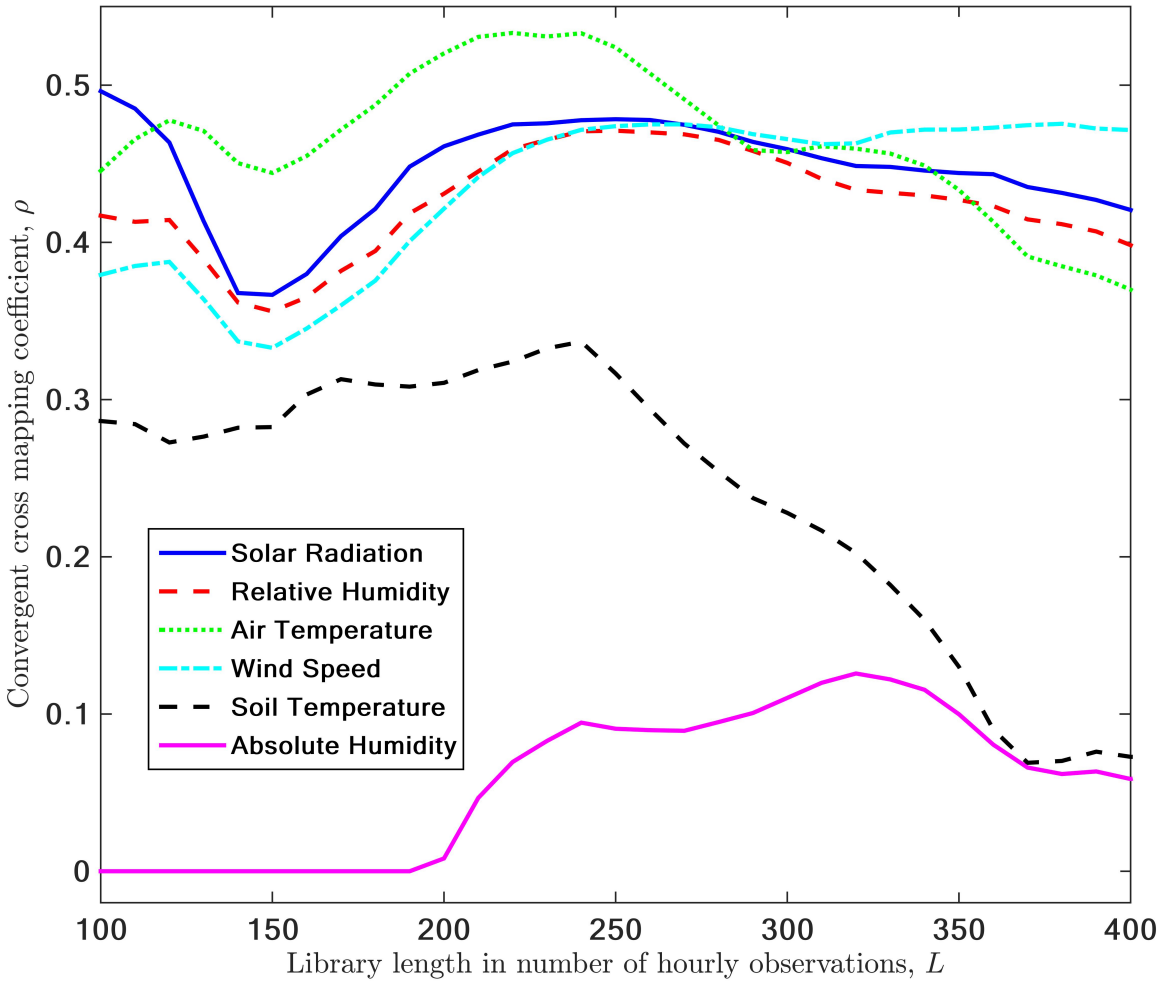

Fig. S1 The CCM coefficient, ρ, between driver signals (meteorological variables) and response signals (spore concentration) for the 2012 monitoring period. These figures identify solar radiation, relative humidity, air temperature, and wind speed as the most important controlling signals with the bifurcation identifying the meteorological conditions that have the greatest influence on spore concentrations.

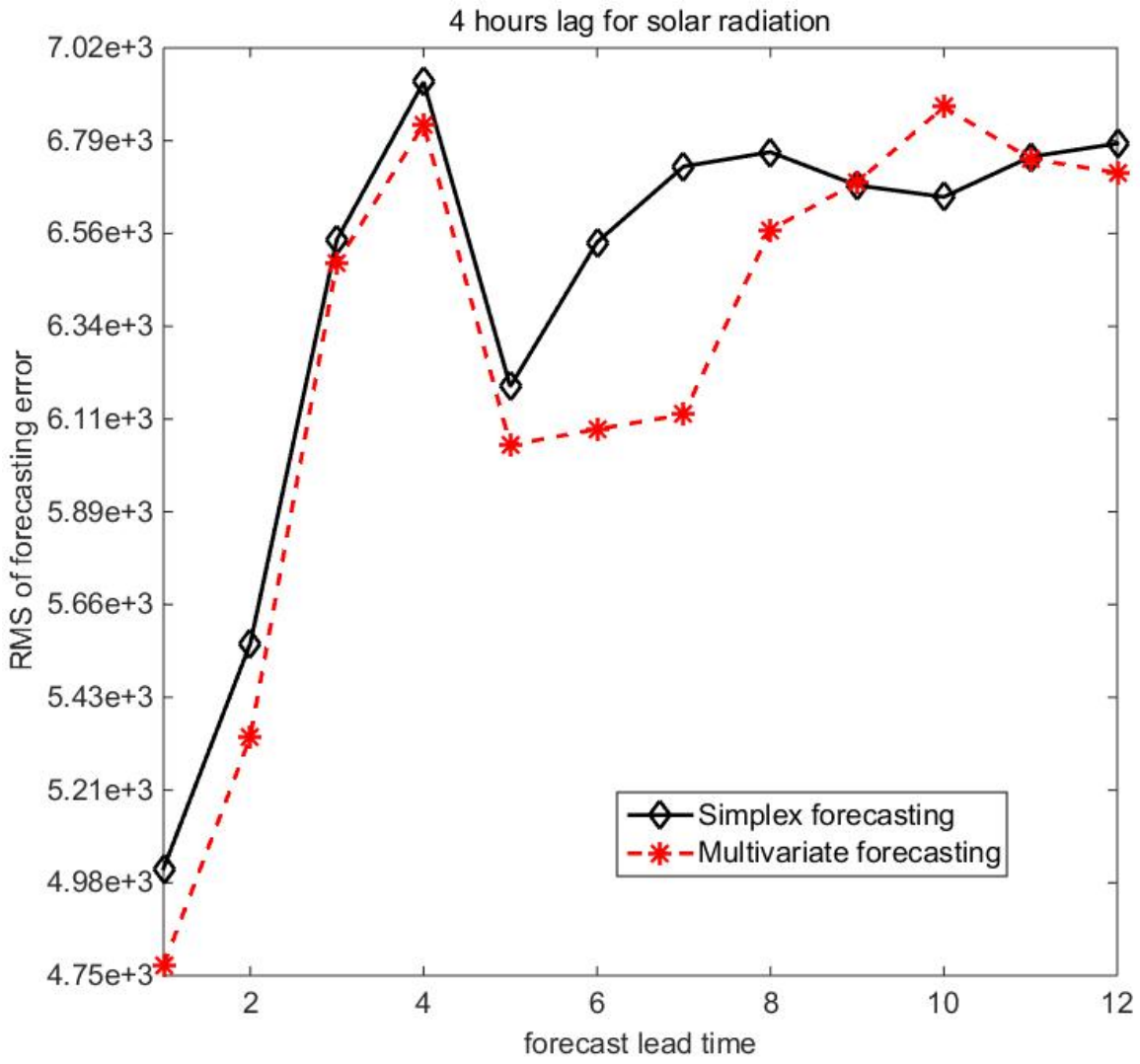

Fig. S2 Forecast root mean square (RMS) errors in cases of multivariate forecasting, solar radiation as the augmented information (dashed red line), and single variable forecasting (solid black line). A lead time of 4 hours is considered between the solar radiation signal and the response signal. Augmentation of the solar radiation improves the RMS of errors for all forecast lead times except for a 10-hour forecast lead time.

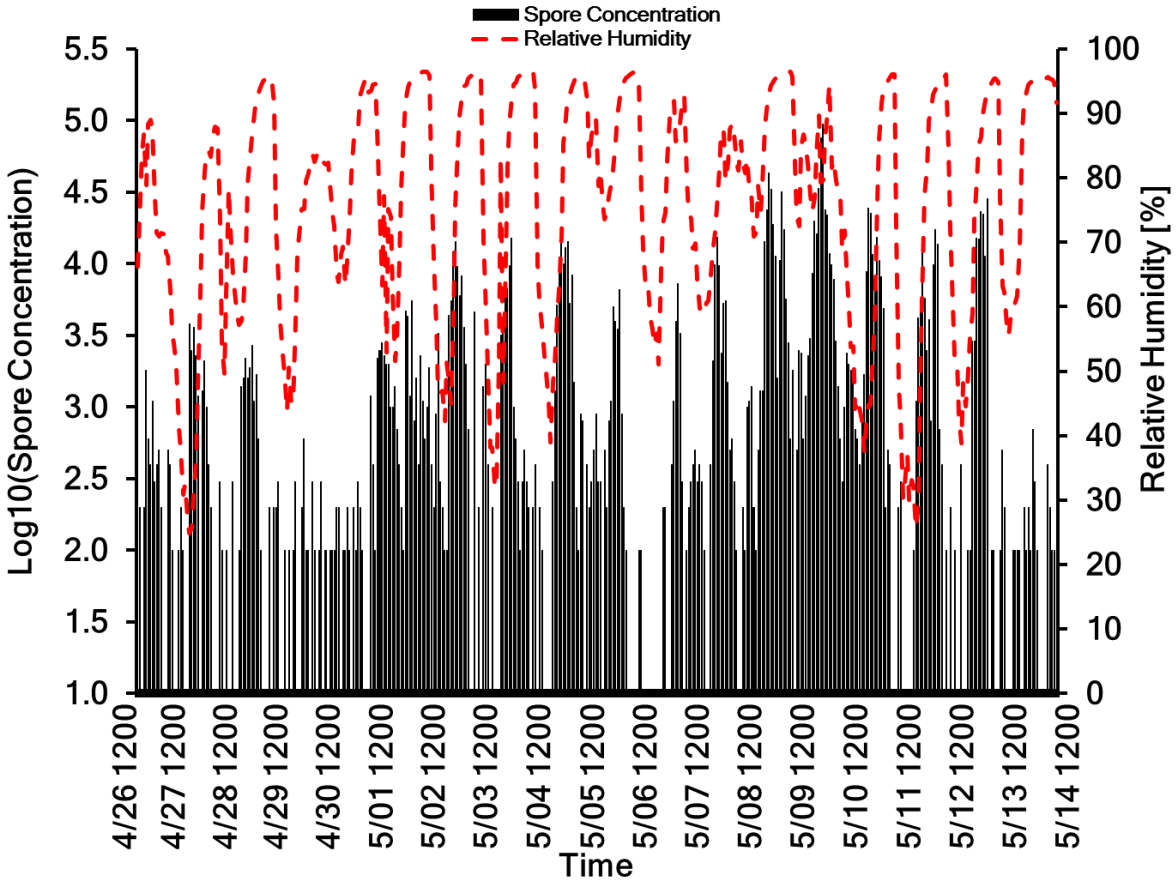

Fig. S3 Hourly spore concentration (black bar graph) and relative humidity (dashed red line) for a field source of *F. graminearum* between 1800 hours 26 April 2012 to 1100 hours 14 May 2012. A single strain of *F. graminearum* (FGVA4) was introduced within a 1-acre wheat field monitored for airborne spore concentration using a Quest volumetric spore sampler.

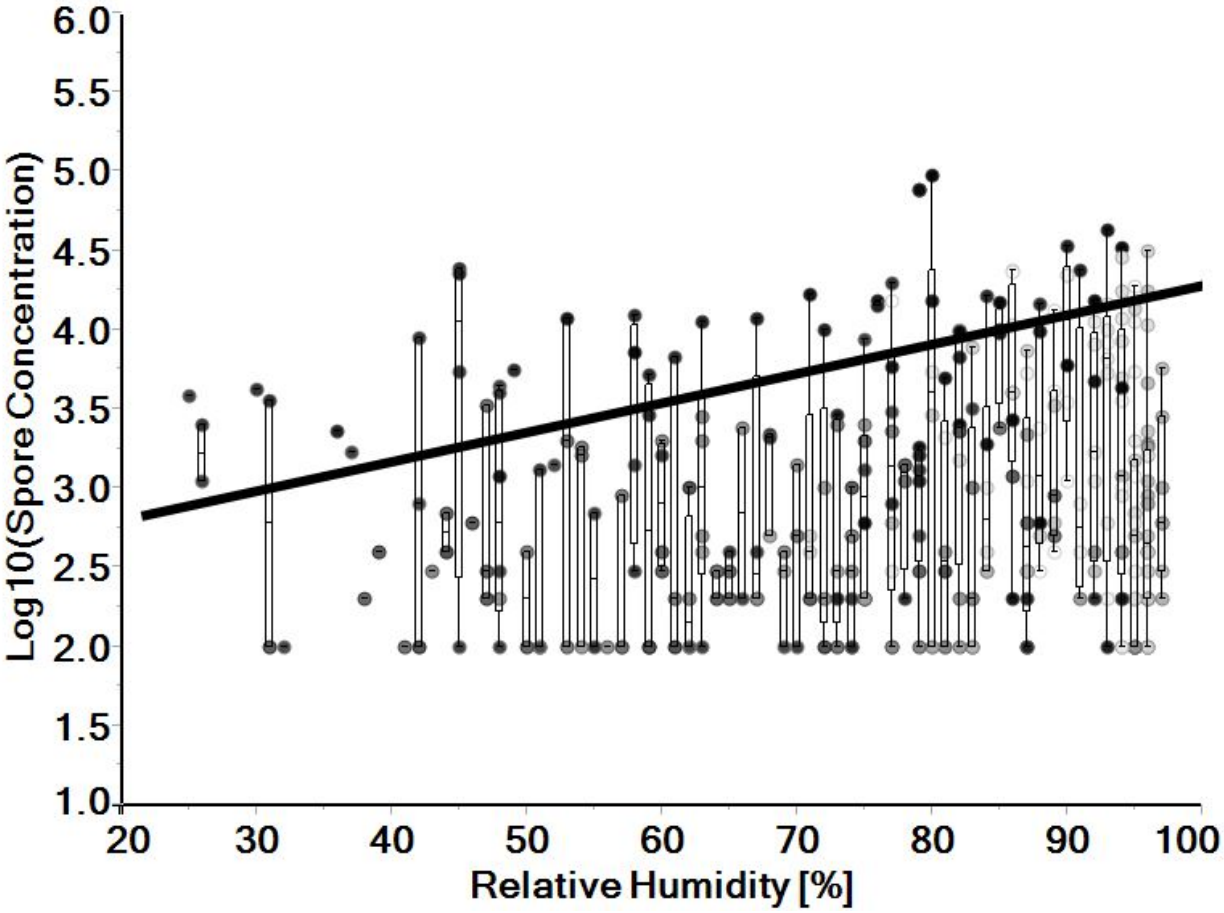

Fig. S4 Spore concentration and the relationship to relative humidity (greyscale markers) for a field-scale source in 2012. The shading distinguishes nighttime (black and white) from daytime (grey variants) events. The black line fits the highest values of spore concentration at each 1% range in relative humidity to illustrate where an apparent threshold occurs.

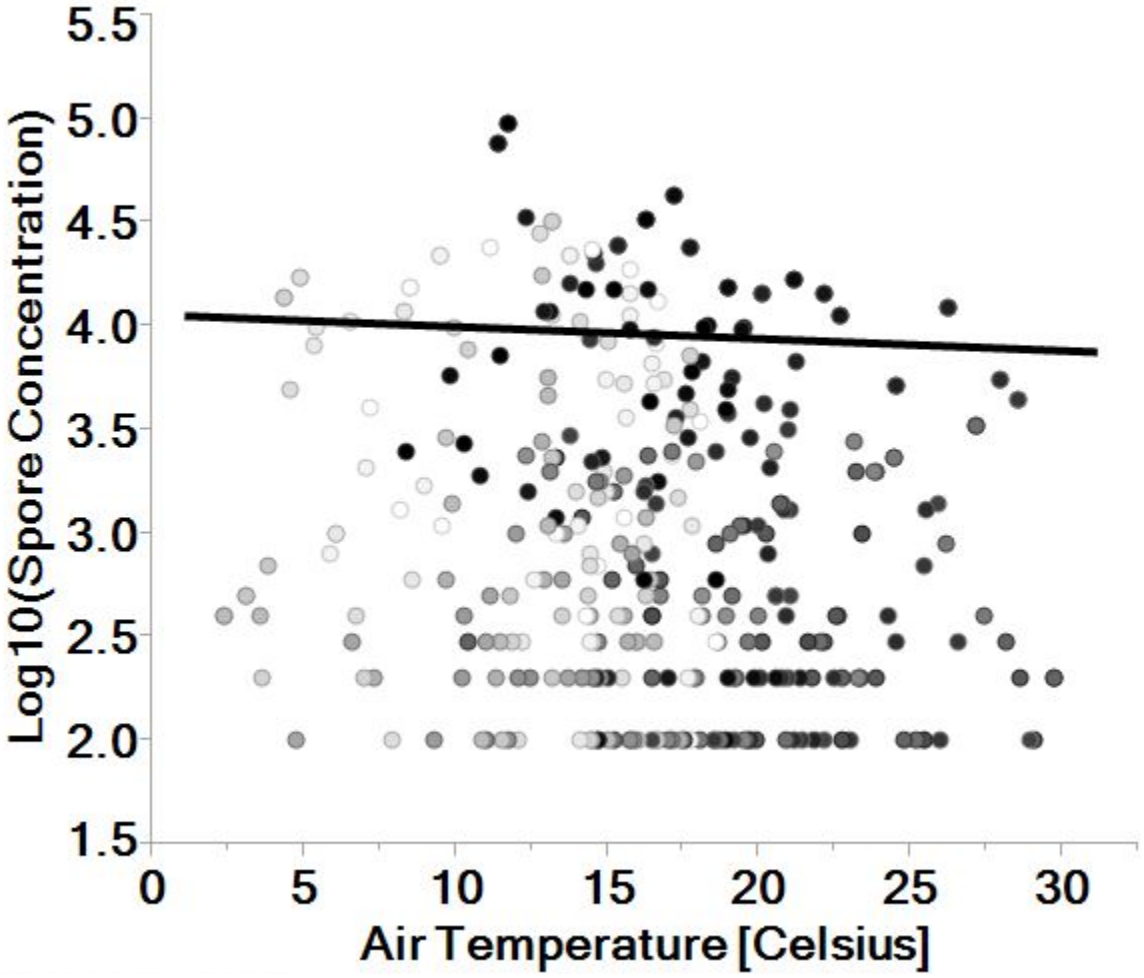

Fig. S5 Spore concentration versus air temperature for a field-scale source of *F. graminearum* during the monitoring period in 2012. The shading distinguishes nighttime (black and white) from daytime (grey variants) events. The black line fits the highest values of spore concentration for each 1° C range in temperature to illustrate where an apparent threshold occurs.

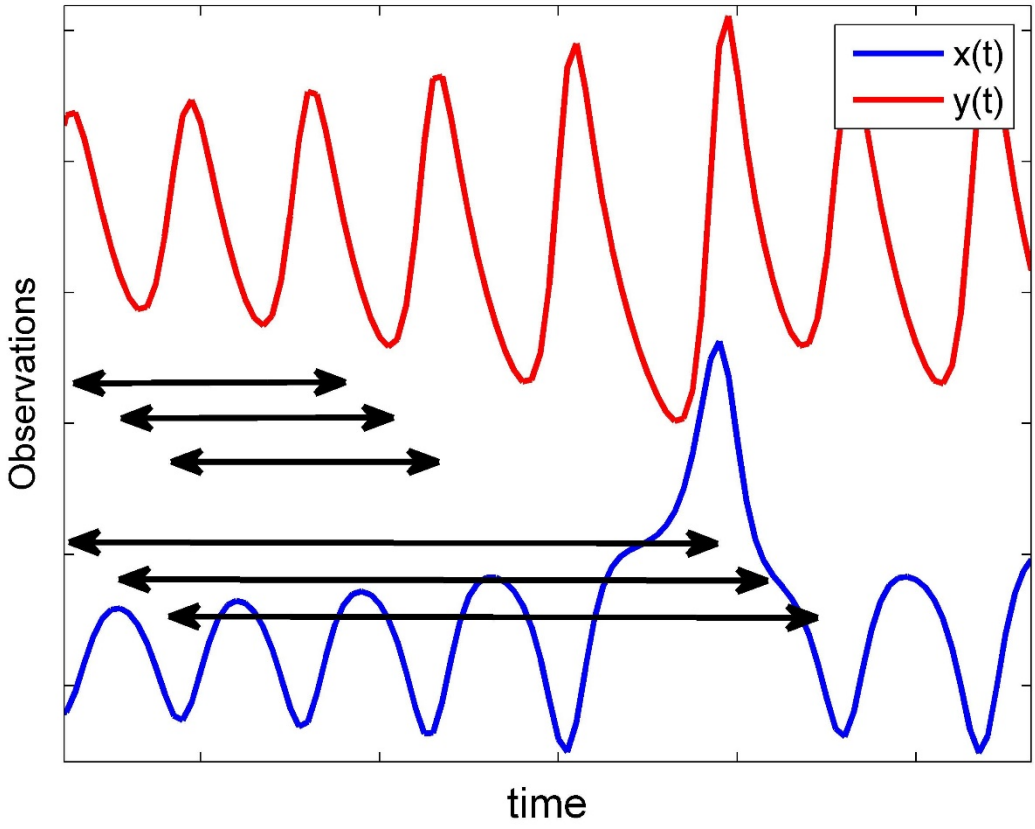

Fig. S6 Schematic of two time series indicating concept of library length**.** An illustration indicating two time-series of observations and three sets of libraries. For a given library length, one sweeps along the full time-series considering all possible sub-time-series, as shown schematically. Thus, for longer library lengths, there will be fewer possible sub-time-series.

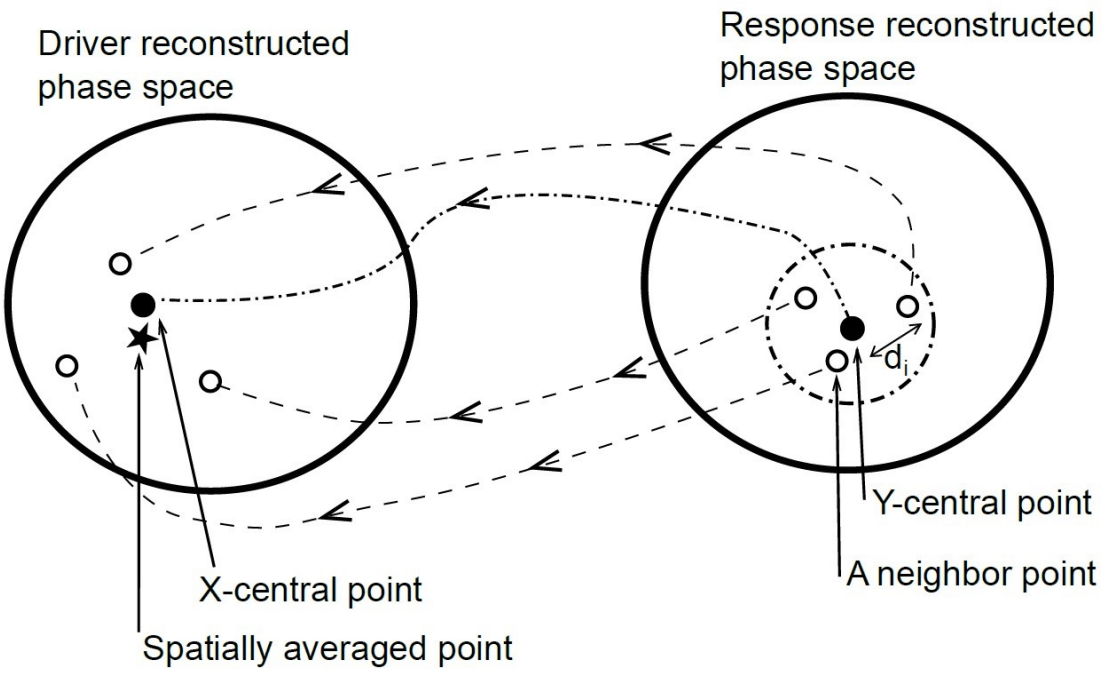

Fig. S7 Schematic of the reconstructed phase spaces of two variables and the process for calculation of ρ. For each E-dimensional Y-central point (black filled circle, right) in the response reconstructed phase space, sufficient numbers of nearest neighbor points are selected (circles, right) and their distance, d_i, to the Y-central point is determined. For each neighbor point, its contemporaneous point in the driver reconstructed phase space is determined (circles, left). These points are weighted by d_i's and averaged. The CCM coefficient is defined as the correlation between the X-central (black filled circle, left) and the recovered (star, left) points.

Table S1 Results of t-test between RMS of errors corresponding to the cases with and without augmented environmental signals.

Parameter	**Analysis between RMS of errors*†ᵛ**								
Delay (hours)	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Solar radiation	**+1**	**+1**	**+1**	**0**	**+1**	**+1**	**+1**	**+1**	**0**
Relative humidity	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**+1**	**0**	**0**
Air temperature	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Wind speed	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Soil temperature	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Absolute humidity	**-1**	**-1**	**0**	**-1**	**-1**	**-1**	**-1**	**0**	**-1**

The analysis includes cases of augmentation of solar radiation, relative humidity, air temperature, wind speed, soil temperature and absolute humidity with the spore concentration signal.

* 0 indicates cases that fail to reject the hypothesis

† +1 indicates cases that reject the null hypothesis and improve the forecast

ᵛ -1 indicates cases that reject the null hypothesis and do not improve the forecast

Table S2 Results of bivariate analysis of spore concentration and meteorological variables.

Year	Indep. term	Regression equation to predict spore concentrations[γαδ]	R^2 (*P*-value)
2011	**Wind speed[θ]**	**$\log_{10}$(Concentration) = 3.28 – 0.32*WS**	**0.12 (< 0.0001)**
		$\log_{10}$(Concentration) = 4.95 – 0.71*WS	**0.55 (< 0.0001)**
	Air temperature[θ]	**$\log_{10}$(Concentration) = 4.49 - 0.086*AT**	**0.21 (<0.0001)**
		$\log_{10}$(Concentration) = 6.38 - 0.12*AT	**0.59 (<0.0001)**
	Relative humidity[θ]	**$\log_{10}$(Concentration) = 1.00 + 0.024*RH**	**0.20 (< 0.0001)**
		$\log_{10}$(Concentration) = -0.015 + 0.049*RH	**0.66 (< 0.0001)**
2012	**Wind speed[θ]**	**$\log_{10}$(Concentration) = 3.14 – 0.13*WS**	**0.047 (< 0.0001)**
		$\log_{10}$(Concentration) = 4.43 – 0.34*WS	**0.46 (< 0.0001)**
	Air temperature[θ]	**$\log_{10}$(Concentration) = 3.37 - 0.025*AT**	**0.030 (0.0020)**
		$\log_{10}$(Concentration) = 4.06 - 0.0058*AT	**0.0057 (0.70)**
	Relative humidity[θ]	**$\log_{10}$(Concentration) = 2.58 + 0.0050*RH**	**0.014 (0.034)**
		$\log_{10}$(Concentration) = 2.43 + 0.018*RH	**0.29 (< 0.0001)**

Relationships between independent model parameters versus spore concentrations for 2011 and 2012 monitoring periods. Each equation represents the linear fit predicting the concentrations of spores per cubic meter based upon the independent variables of wind speed, (m s^{-1}), air temperature (^{o}C), and relative humidity (%).

[γ] WS is wind speed in m s^{-1}

[α] AT is air temperature in ^{o}C

[δ] RH is relative humidity in percent

[θ] First row is linear fit of entire data set and second row is linear fit of threshold line