SUPPLEMENTARY MATERIAL

Supplementary Methods

Establishing the SARIMA Usual ARIMA is specified by three components: the autoregressive (AR) component, the integrated (I) component, and the moving average (MA) component[1], which involves predicting future epidemics based on a non-seasonal time series. The SF incidence frequently has notable seasonal effects[2], and hence a seasonal ARIMA (SARIMA) should be adopted, it is an extension of the ARIMA by including the seasonal versions of the three components above, which is designed to capture the underlying patterns and trends by considering both the seasonal and non-seasonal components of a series[1]. The SARIMA is usually denoted as SARIMA(p, d, q)(P, D, Q)$_S$, where p refers to the number of AR terms, d represents the degree of differencing, q signifies the number of MA terms, P, D, and Q stand for the seasonal terms above (i.e. SAR, SMA, and seasonal difference), and S is the number of periods in a season. SARIMA requires determining the six parameters above through four steps. First, SARIMA assumes stationarity, and thus stationarity of the SF incidence series was analyzed by a KPSS unit root test[3], this statistic rejects stationarity when $P < 0.05$, indicating that differencing was required to achieve data stationarity, and otherwise not. Second, identifying the appropriate structure by inspecting the autocorrelation function (ACF) and partial ACF (PACF) plots that help roughly determine the values of p, q, P, and Q[4,5]. A series of combinations emerged thereof, the best one was identified by maximizing the log-likelihood (LL), and minimizing the Akaike's information criterion (AIC), corrected AIC (CAIC), and Bayesian information criterion (BIC)[1]. Third, conducting model diagnostics to judge whether the resulting residuals were white noise based on the Ljung-Box Q test, autocorrelogram, and partial autocorrelogram[1,5]. Finally, once the best model satisfied the required tests, it could be determined for forecasting purposes.

Establishing the SARFIMA Time series often has a complex interplay between observed values, which is characterized by a gradual decrease in magnitude over time, following a hyperbolic decay (HD) pattern[6]. Unlike SARIMA which assumes that the autocorrelation decays exponentially, SARFIMA allows for a HD of autocorrelation, thus accommodating long-range dependence, which has been the most commonly used model to analyze the underlying mechanisms driving HD[6]. By incorporating the fractional integration (d_f), SARFIMA provides a flexible framework for capturing both short and long memory simultaneously[6,7]. d_f in different ranges suggests various features of a series[8]. Usually, the range $d_f \in (-1, 0.5)$ is used, so if $d_f \in (-0.5, 0)$, indicating the invertibility of the series; if $d_f \in (-1, -0.5)$, indicating the anti-persistence of the series; if $d_f = 0$, indicating the short memory and mean-reverting process of the series; and if $d_f \in (0, 0.5)$, indicating the long-range persistence of the series[6,8]. Often, a SARFIMA is denoted as SARFIMA (p, d^*, q)(P, D^*, Q)$_S$, where $d^* = d + d_f$ and $D^* = D + D_f$, d_f or D_f represents the fractional integration, and d or D signifies the integer part (where d or $D \geq 0$)[6]. The Hurst (H) exponent serves as a valuable statistical measure used to analyze the long-term memory and predictability of a time series as it quantifies the degree of persistence or anti-persistence present in a series[9]. The relationship between H and d_f is denoted as d_f or $D_f = H - 0.5$, so if $H > 0.5$, indicating a persistent series; if $H < 0.5$, suggesting an anti-persistent series; and if $H = 0.5$, showing a random walk of series[6,7]. The computation of H includes some techniques such as rescaled range (R/S) analysis or detrended fluctuation analysis[9]. This study used the corrected R/S to determine whether the SF incidence series displays long-range properties. Constructing the SARFIMA requires selecting the best modes as it is under the assumption of multiple modes (i.e. beginning the fits with multiple starting values), causing more than one mode[6]. The best one was identified by maximizing LL and minimizing AIC and BIC[6,8]. The other steps required estimating the parameters and conducting model diagnostics, which followed what was indicated in SARIMA.

REFERENCES

1. Hyndman RJ, Khandakar Y. Automatic Time Series Forecasting: The forecast Package for R. Journal of statistical software, 2008; 27, 1−22.
2. Lamagni T, Guy R, Chand M, et al. Resurgence of scarlet fever in England, 2014-16: a population-based surveillance study. Lancet Infect Dis, 2018; 18, 180−7.
3. Kwiatkowski D, Phillips PCB, Schmidt P, et al. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. Journal of Econometrics, 1992; 54, 159−78.
4. Kuan MM. Applying SARIMA, ETS, and hybrid models for prediction of tuberculosis incidence rate in Taiwan. PeerJ, 2022; 10, e13117.
5. Wu WW, Li Q, Tian DC, et al. Forecasting the monthly incidence of scarlet fever in Chongqing, China using the SARIMA model.

Epidemiol Infect, 2022; 150, e90.
6. Veenstra J. Persistence and Anti-persistence: Theory and Software (2013). Electronic Thesis and Dissertation Repository, The University of Western Ontario, 2013.
7. Qi C, Zhang D, Zhu Y, et al. SARFIMA model prediction for infectious diseases: application to hemorrhagic fever with renal syndrome and comparing with SARIMA. BMC Medical Research Methodology, 2020; 20, 243.
8. Chang F, Huang H, Chan A H S, et al. Capturing long-memory properties in road fatality rate series by an autoregressive fractionally integrated moving average model with generalized autoregressive conditional heteroscedasticity: A case study of Florida, the United States, 1975-2018. J Safety Res, 2022; 81, 216−24.
9. ANNIS AA, LLOYD EH. The expected value of the adjusted rescaled Hurst range of independent normal summands. Biometrika, 1976; 63, 111−6.

Supplementary Table S1. Resultant candidate modes under the SARFIMA(3, 0, 1)(3, −0.347, 0)$_{12}$

Modes	AIC	BIC	LL
Mode 1	1654.925	1690.048	−816.463
Mode 2	1656.486	1691.608	−817.243
Mode 3	1656.844	1691.967	−817.422
Mode 4	1658.588	1693.71	−818.294
Mode 5	1658.738	1693.861	−818.369
Mode 6	1659.058	1694.18	−818.529
Mode 7	1659.949	1695.072	−818.975
Mode 8	1660.268	1695.39	−819.134
Mode 9	1662.806	1697.928	−820.403
Mode 10	1663.734	1698.856	−820.867
Mode 11	1665.047	1700.17	−821.524
Mode 12	1666.785	1701.907	−822.392
Mode 13	1666.866	1701.989	−822.433
Mode 14	1667.475	1702.597	−822.737
Mode 15	1672.506	1707.629	−825.253
Mode 16	1675.773	1710.896	−826.887
Mode 17	1683.103	1718.225	−830.551
Mode 18	1692.033	1727.156	−835.017
Mode 19	1692.24	1727.362	−835.12
Mode 20	1692.442	1727.565	−835.221
Mode 21	1700.832	1735.954	−839.416
Mode 22	1702.176	1737.299	−840.088
Mode 23	1703.389	1738.512	−840.695
Mode 24	1714.219	1749.341	−846.109
Mode 25	1714.768	1749.89	−846.384
Mode 26	1732.879	1768.002	−855.44
Mode 27	1738.179	1773.301	−858.089

***Note*.** SARFIMA, seasonal autoregressive fractionally integrated moving average; AIC, Akaike's information criterion; BIC, Bayesian information criterion; LL, log−likelihood.

Supplementary Table S2. Resultant candidate modes under the SARFIMA(2, −0.302, 1)(1, 0.471, 2)$_{12}$

Modes	AIC	BIC	LL
Mode 1	1561.067	1595.431	−769.534
Mode 2	1561.444	1595.808	−769.722
Mode 3	1561.448	1595.811	−769.724
Mode 4	1561.448	1595.811	−769.724
Mode 5	1561.45	1595.814	−769.725
Mode 6	1561.486	1595.85	−769.743
Mode 7	1561.495	1595.859	−769.748
Mode 8	1561.499	1595.863	−769.75
Mode 9	1561.502	1595.865	−769.751
Mode 10	1561.506	1595.87	−769.753
Mode 11	1561.509	1595.873	−769.755
Mode 12	1561.511	1595.874	−769.755
Mode 13	1561.511	1595.875	−769.755
Mode 14	1561.512	1595.875	−769.756
Mode 15	1561.512	1595.875	−769.756
Mode 16	1561.512	1595.876	−769.756
Mode 17	1561.512	1595.876	−769.756
Mode 18	1561.514	1595.878	−769.757
Mode 19	1561.74	1596.103	−769.87
Mode 20	1561.753	1596.117	−769.877
Mode 21	1561.776	1596.14	−769.888
Mode 22	1561.791	1596.154	−769.895
Mode 23	1561.811	1596.174	−769.905
Mode 24	1561.83	1596.194	−769.915
Mode 25	1561.843	1596.207	−769.922
Mode 26	1561.846	1596.21	−769.923
Mode 27	1561.855	1596.218	−769.927
Mode 28	1561.86	1596.224	−769.93
Mode 29	1561.865	1596.229	−769.933
Mode 30	1561.867	1596.231	−769.934
Mode 31	1561.883	1596.247	−769.941
Mode 32	1561.968	1596.331	−769.984
Mode 33	1562.251	1596.615	−770.126
Mode 34	1562.261	1596.625	−770.131
Mode 35	1562.871	1597.235	−770.436
Mode 36	1563.323	1597.687	−770.661
Mode 37	1563.362	1597.726	−770.681
Mode 38	1563.447	1597.81	−770.723
Mode 39	1563.761	1598.124	−770.88
Mode 40	1564.014	1598.378	−771.007
Mode 41	1564.144	1598.507	−771.072

Continued

Modes	AIC	BIC	LL
Mode 42	1564.149	1598.512	−771.074
Mode 43	1564.15	1598.513	−771.075
Mode 44	1564.158	1598.521	−771.079
Mode 45	1564.159	1598.523	−771.08
Mode 46	1564.162	1598.526	−771.081
Mode 47	1564.189	1598.553	−771.095
Mode 48	1564.213	1598.576	−771.106
Mode 49	1564.377	1598.741	−771.189
Mode 50	1564.683	1599.047	−771.342
Mode 51	1564.755	1599.119	−771.377
Mode 52	1564.992	1599.356	−771.496
Mode 53	1565.249	1599.612	−771.624
Mode 54	1565.656	1600.02	−771.828
Mode 55	1566.507	1600.87	−772.253
Mode 56	1567.261	1601.625	−772.631
Mode 57	1567.358	1601.722	−772.679
Mode 58	1567.595	1601.959	−772.798
Mode 59	1567.683	1602.047	−772.842
Mode 60	1567.746	1602.109	−772.873
Mode 61	1568.846	1603.21	−773.423
Mode 62	1569.285	1603.649	−773.643
Mode 63	1570.262	1604.625	−774.131
Mode 64	1570.905	1605.268	−774.452
Mode 65	1570.961	1605.325	−774.481
Mode 66	1571.379	1605.742	−774.689
Mode 67	1573.295	1607.658	−775.647
Mode 68	1574.893	1609.257	−776.447
Mode 69	1575.204	1609.567	−776.602
Mode 70	1575.716	1610.079	−776.858
Mode 71	1576.168	1610.531	−777.084
Mode 72	1576.184	1610.547	−777.092
Mode 73	1576.558	1610.921	−777.279
Mode 74	1576.932	1611.296	−777.466
Mode 75	1580.158	1614.521	−779.079
Mode 76	1590.366	1624.73	−784.183
Mode 77	1590.854	1625.218	−784.427
Mode 78	1595.913	1630.276	−786.956
Mode 79	1598.18	1632.544	−788.09
Mode 80	1600.319	1634.682	−789.159
Mode 81	1611.598	1645.962	−794.799

***Note*.** SARFIMA, seasonal autoregressive fractionally integrated moving average; AIC, Akaike's information criterion; BIC, Bayesian information criterion; LL, log-likelihood.

Supplementary Table S3. Forecasts between January 2018 and December 2019 from the SARIMA and SARFIMA

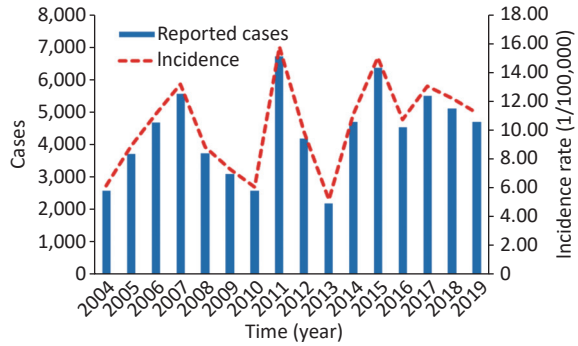
Time	Observations	SARIMA(2, 1, 2)(1, 1, 2)$_{12}$		SARFIMA(2, -0.302, 1)(1, 0.471, 2)$_{12}$	
		Forecasts	95% *CI*	Forecasts	95% *CI*
18-January	638	577	408 to 746	654	471 to 837
18-February	214	244	30 to 457	340	72 to 609
18-March	282	328	98 to 558	423	133 to 712
18-April	401	393	153 to 633	454	154 to 754
18-May	797	764	510 to 1018	654	347 to 961
18-June	855	856	575 to 1137	713	401 to 1,024
18-July	353	571	253 to 889	441	126 to 756
18-August	99	305	−48 to 658	219	−98 to 536
18-September	138	305	−70 to 681	266	−53 to 585
18-October	245	380	−9 to 769	347	27 to 667
18-November	472	676	279 to 1,074	605	284 to 926
18-December	631	875	467 to 1,283	709	387 to 1,030
19-January	450	598	170 to 1,026	453	127 to 779
19-February	182	349	−102 to 801	208	−122 to 538
19-March	343	450	−24 to 925	308	−24 to 639
19-April	418	442	−51 to 934	361	28 to 693
19-May	538	730	225 to 1,235	554	221 to 887
19-June	587	774	258 to 1,289	607	273 to 940
19-July	380	491	−36 to 1,018	365	31 to 699
19-August	119	287	−253 to 828	164	−170 to 498
19-September	217	360	−197 to 918	216	−118 to 550
19-October	300	454	−120 to 1,029	297	−37 to 631
19-November	559	734	145 to 1,322	545	211 to 879
19-December	617	883	283 to 1,484	629	295 to 963

***Note*.** SARIMA, seasonal autoregressive integrated moving average; SARFIMA, seasonal autoregressive fractionally integrated moving average; *CI*, confidence interval.

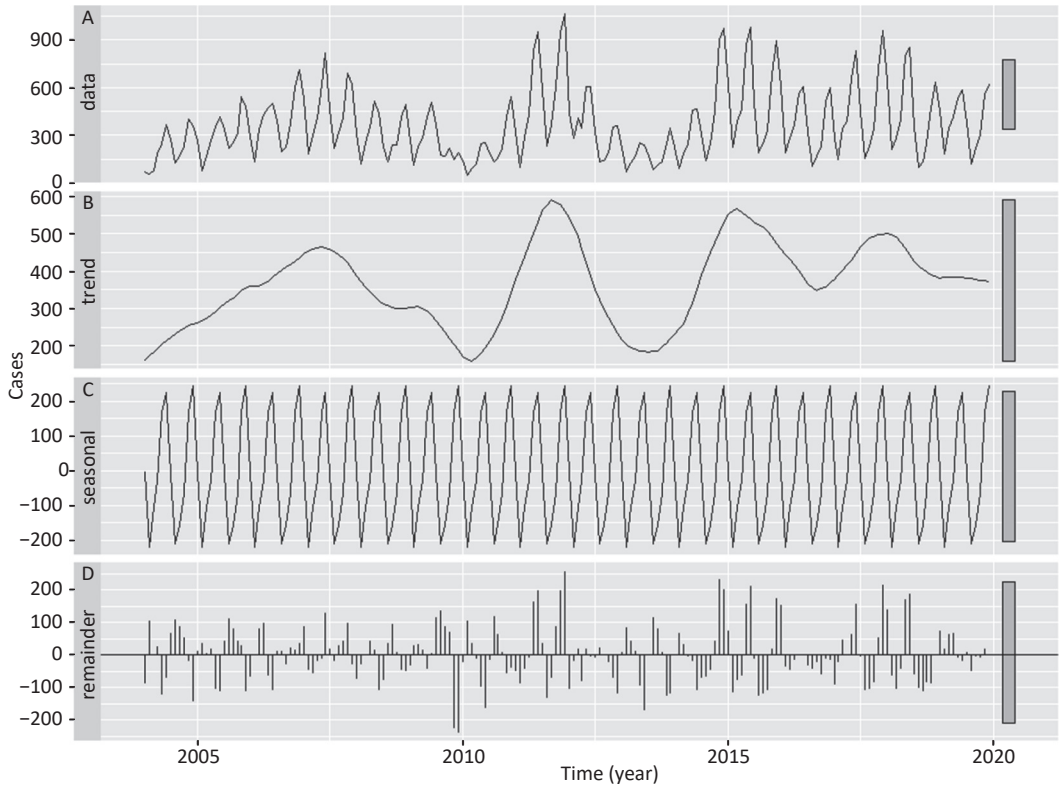
Supplementary Table S4. Identified possible SARIMA with the AIC, CAIC, BIC, and LL values

Models	AIC	CAIC	BIC	LL	Ljung-Box Q test	
					χ^2	*P*
SARIMA(3, 0, 1)(0, 1, 1)$_{12}$	2012.96	2013.48	2031.70	−1000.48	0.108	0.743
SARIMA(3, 0, 1)(3, 1, 0)$_{12}$	2007.61	2008.51	2032.60	−995.80	0.047	0.829
SARIMA(3, 0, 1)(2, 1, 0)$_{12}$	2018.84	2019.54	2040.71	−1002.71	0.130	0.718
SARIMA(3, 0, 1)(1, 1, 0)$_{12}$	2050.74	2051.26	2069.49	−1019.37	0.053	0.818
SARIMA(3, 0, 1)(1, 1, 1)$_{12}$	2014.59	2015.29	2036.46	−1000.30	0.120	0.729
SARIMA(2, 0, 1)(3, 1, 0)$_{12}$	2014.53	2015.23	2036.40	−1000.27	0.004	0.949
SARIMA(1, 0, 1)(3, 1, 0)$_{12}$	2014.55	2015.07	2033.29	−1001.27	0.114	0.735

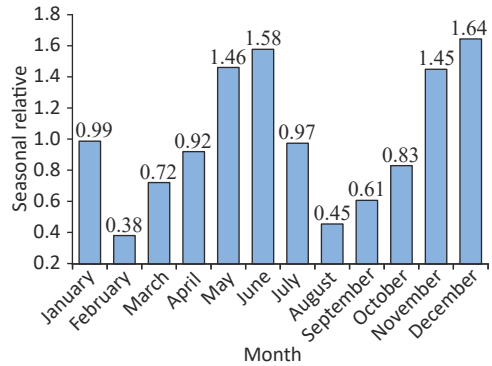
***Note*.** SARIMA, seasonal autoregressive integrated moving average; AIC, Akaike's information criterion; CAIC, corrected Akaike's information criterion; BIC, Bayesian information criterion; LL, log-likelihood.



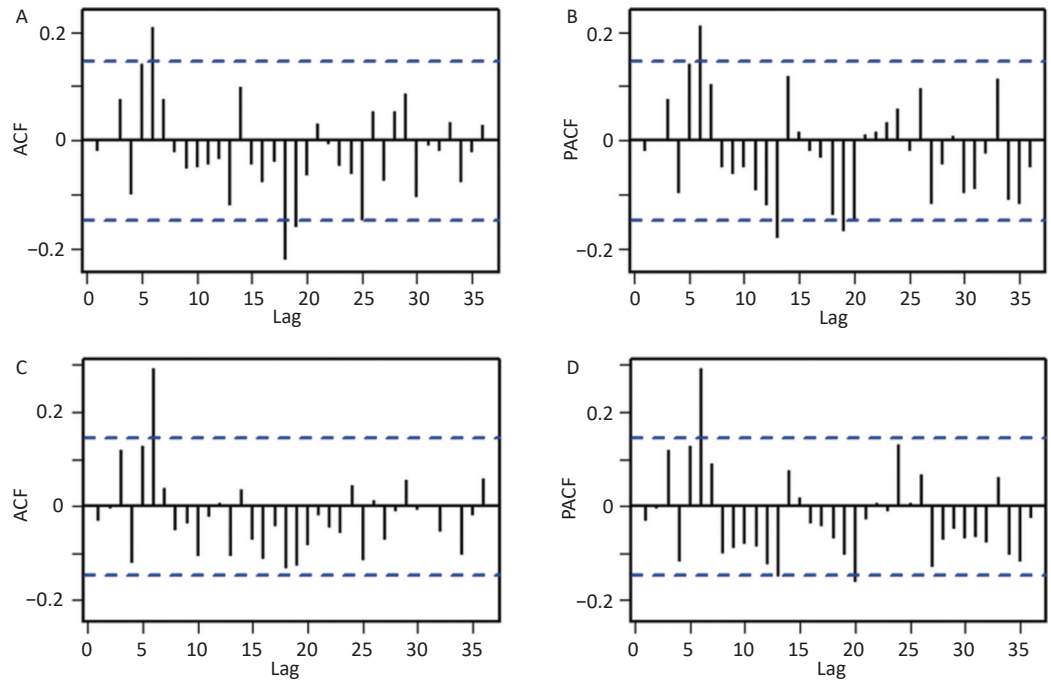
Supplementary Figure S1. Yearly incident cases and incidence rate in Liaoning during 2004–2019. This plot pinpoints that the SF outbreak occurred in 2011 and there is a periodic cycle pattern of around 4–7 years.



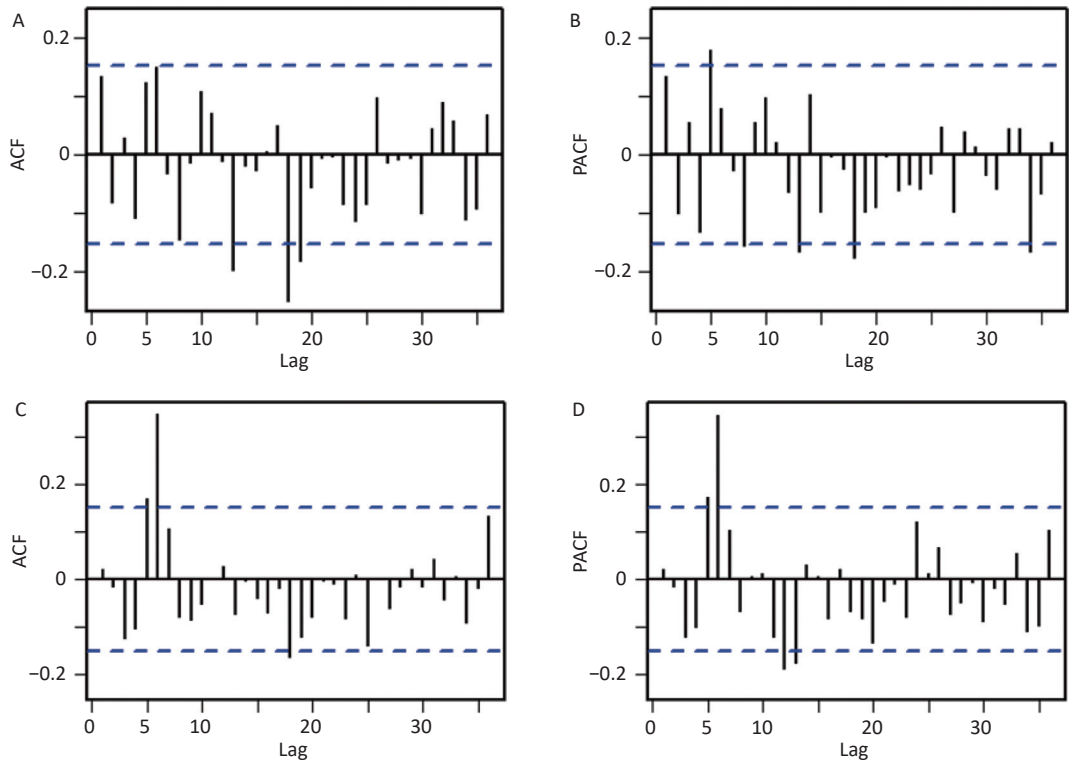
Supplementary Figure S2. A seasonal decomposition of the SF incidence series based on the STL technique. The (A) SF series is decomposed into (B) seasonal, (C) trend, and (D) irregular parts. It seems that there is a periodic outbreak pattern and a clear seasonality in SF incidence.



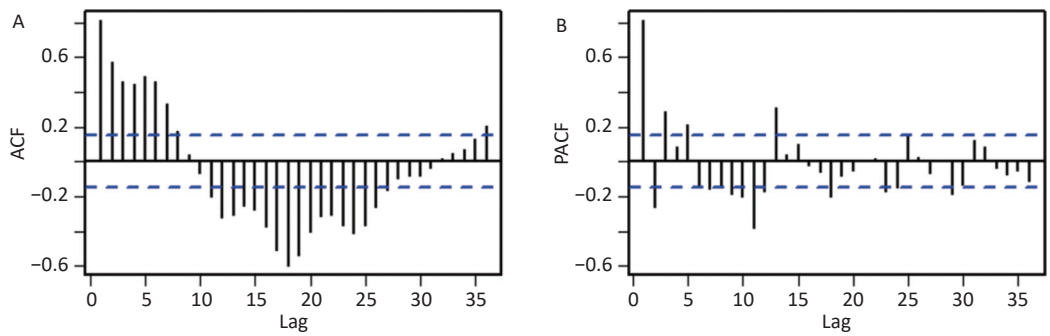
Supplementary Figure S3. The decomposed seasonal relative (SR) for the SF morbidity series using the multiplicative decomposition method. A value of SR = 1 means that the incidence for that period is exactly the same as the average. A value of SR > 1 means the incidence is higher than the average (indicating a high-risk season), and a value of SR < 1 means this period's incidence is lower than the average (indicating a low-risk season). As shown, SF epidemics present pronounced dual seasonal patterns per year.



Supplementary Figure S4. ACF and PACF plots for the residual series under the SARIMA and SARFIMA. (A) ACF and (B) PACF plots for the residual series under the SARIMA, (C) ACF and (D) PACF plots for the residual series under the SARFIMA. Here the correlogram demonstrates that most spikes fall within the 95% *CI* except for few outside this significance bounds (which is also reasonable because some high-order correlations easily exceed the significance bounds by chance alone), indicating that there is little evidence of non-white noise in the forecast errors.



Supplementary Figure S5. ACF and PACF plots for the residual series under the SARIMA and SARFIMA based on the data during 2004-2017 in Liaoning. (A) ACF and (B) PACF plots for the residual series under the SARIMA, (C) ACF and (D) PACF plots for the residual series under the SARFIMA. Here the correlogram demonstrates that most spikes fall within the 95% *CI* except for few outside this significance bounds (which is also reasonable because some high-order correlations easily exceed the significance bounds by chance alone), indicating that there is little evidence of non-white noise in the forecast errors.



Supplementary Figure S6. ACF and PACF plots for the seasonally differenced series in Liaoning. (A) ACF plot, and (B) PACF plot. The significant spike at lag 3 in the PACF indicates that the maximum orders may be 3 in the non-seasonal AR component, the significant spike at lag 10, 11, and 12, along with 23, 24, and 25 in the ACF suggests that the maximum orders may be 3 in the seasonal AR component. The significant spikes at lag 12, 24, and 36 in the ACF suggests that the maximum orders may be 1 in the seasonal MA component.