The Rise of West Nile Virus in Southern and Southeastern Europe

Matthew J. Watts et al., (2021)

Extended data extraction and processing methods

Aggregation

All data were aggregated annually to produce the final yearly panel data-set and aggregated at the NUTS 3 country subdivision level, apart from central government spending data which was sourced at the country level. All spatial information was captured at the NUTS 3 level using shapefiles (polygons) sourced from R's 'eurostat' package ((Lahti et al., 2017)).

The Nomenclature of territorial units for statistics (NUTS) is a classification system used to divide economic territories of the EU into three hierarchical sub categories for the purpose of data collection and and statistical analysis:

- NUTS 1: Major socio-economic regions with a population ranging from 3 to 7 million.
- NUTS 2: Basic regions are generally used for the application of regional policies with a population ranging from 800,000 to 3 million.
- NUTS 3: Small regions for specific diagnoses with a population ranging from 150,000 to 800,000.

For further details see https://ec.europa.eu/eurostat/web/nuts/background.

WNF Case Data

WNV case data were provided at request by the European Centre for Disease Prevention and Control (www.ecdc.europa.eu). Case data are collected weekly by EU member states and affiliates. Data were aggregated at NUTS 3 country subdivisions ((Eurostat, 2020)). Positive cases were confirmed by at least one of the following techniques: 1); isolating WNV or WNV nucleic acid from blood or cerebrospinal fluid (CSF); 2) inducing a WNV-specific antibody response (either IgG / IgM) in a serological test. All cases were aggregated yearly to create the annual panel data-set.

Economic, Socio-Economic and Demographic Factors

Economic data were extracted from the Eurostat database (https://ec.europa.eu/eurostat/data/database), which provides comparable statistics and indicators and is presented in yearly time series. To capture factors determining the economic crisis, austerity and cuts to public spending we selected NUTS3 regional level Gross Domestic Product (GDP); and country level agriculture, forestry, fisheries spending, waste water spending Health spending. The "Agriculture, forestry, fisheries spending" variable captures spending in rural areas that help to improve the environment and agricultural development, that can benefit agricultural workers and/or mechanize production ((Eurostat, 2019)). In order to represent spending before and after the economic crisis, we created a baseline index for each variable set at 2007 levels, which represented negative or positive growth from the point just before the economic crisis hit Europe.

Climate Data

Climate data were sourced from the E-OBS Gridded Data-set ((Cornes et al., 2018)). This data-set was created using a series of daily temperature and rainfall observations at meteorological stations throughout Europe. R's "Raster Extract" function from the "Raster" package ((van Etten, 2012)) was used to extract and aggregate cell values to each NUTS 3 region. The subsequent regional values were then processed further to create regional seasonal variables: “Mean temp winter (°C)”, “Mean temp spring (°C)”, “Mean temp summer (°C)”, “Days of rain in winter, “Days of rain in spring” and "Days of rain in summer”. Winter was designated as December to March, Spring as March to June, and summer June to September.

Land-use data

Land use statistics were captured at the NUTS3 level using the CORINE Land Cover (CLC) 2006, 2012 and 2018 data-sets ((EU, 2018)). These data-sets provide information on the biophysical characteristics of the Earth's surface in the form of categorical raster data. For each region, we calculated percentage land cover for each of the land-use risk factors identified in our conceptual framework, i.e., “Continuous urban fabric”, “Discontinuous urban fabric', “Wetlands (fresh water)” and “Arable land”. R's SF and Raster packages ((Pebesma, 2018; van Etten, 2012)) were used to extract information for each available year (2006, 2012, 2018). R's Zoo package ((Zeileis and Grothendieck, 2005)) was used to calculate values for missing years, by implementing a linear interpolation method that would predict trends between years, apart from 2019 where 2018 values were used.

Surface Water data

Regional surface water data was sourced using the JRC Monthly Water History, v1.2 data set ((Pekel et al., 2016)) via Google Earth Engine at a 30 meter pixel resolution. This data set contains maps of the location and temporal distribution of surface water from 1984 to 2019 and provides statistics on the extent and change of water surfaces. Data were generated using scenes from Landsat 5, 7, and 8. Each pixel was individually classified into water / non-water using an expert system and the results were collated into a monthly history. Water / non-water count observations were extracted and aggregated by each NUTS 3 region. The sum of the Water / non-water observations were then used to create a % water surface water indicator, which was averaged by season and converted to Z-scores to standardize values. This would help determine if the seasonal water extent was average, below the mean (low), or above the mean (high) for a given year.

Extended Statistical methods

General additive regression model to assess associations of independent variables on WNV case data at regional level

One of the main issues with our data-set is that it did not meet some basic assumptions for statistical inference, and specifically the data are not independent and identically distributed random variables (iid). More specifically, the data-set captured repeated measurements over the same regions, and observations were not independent because of spill over effects from neighboring regions, therefore we needed to implement an appropriate statistical design to control for both temporal and spatial pseudo replication (lack of independence). We could deal with this in two ways, 1) either using a generalized linear mixed model (GLMM) approach, relaxing the assumption of independence and estimating the spatial/temporal correlation between residuals, or 2) model the spatial and temporal dependence in the systematic part of the model ((Aswi et al., 2018)). We opted to use a Generalized Additive Model (GAM) using R's Mgcv statistical package ((Wood, 2011)) because of its versatility and ability to fit complex models that would converge even with low numbers of observations, and could capture potential complex non-linear relationships. One of the advantages of GAMs is that we do not need to determine the functional form of the relationship beforehand. In general, such models transform the mean response to an additive form so that additive components are smooth functions (e.g., splines) of the covariates, in which functions themselves are expressed as basis-function expansions. The spatial auto-correlation in the GAM model was approximated by a Markov random field (MRF) smoother, defined by the geographic areas and their neighborhood structure. We used R's Spdep package ((Bivand et al., 2013)) to create a queen neighbors list (adjacency matrix) based on regions with contiguous boundaries i.e.~those sharing one or more boundary point. We used a full rank MRF, which represented roughly one coefficient for each area. The local Markov property assumes that a region is conditionally independent of all other regions unless regions share a boundary. This feature allowed us to model the correlation between geographical neighbors and smooth over contiguous spatial areas, summarizing the trend of the response variable as a function of the predictors (see section 5.4.2 of ((Wood, 2017)). In order to account for variation in the response variable over time, not attributed to the other explanatory variables in our model, we used a saturated time effect for years, where a separate effect per time point is estimated.

We first tried to fit our model using a Poisson distribution. However, the mean of our dependent variable (WNV cases by region and year) was lower than its variance - E(Y) <Var(Y), suggesting that the data are over-dispersed. We also tried to fit our models using the negative binomial, quasi-Poisson and Tweedie distribution, all particularly suited when the variance is much larger than the mean. After several tests, we concluded that the Tweedie distribution worked well with our data since it can handle excess zeros ((Kurz, 2017)), and allows us to model the incident rate, although results were comparative across all distributions (note that WNV infection count data, offset by a log of population at risk was used for the neg bin and quasi-Poisson models). Analysis of model diagnostic tests did not reveal any major issues; in general residuals appeared to be randomly distributed (see additional information - Figures S10-S11 and Table S1 for diagnostics).

Tweedie distributions are defined as subfamily of (reproductive) exponential dispersion models (ED), with a special mean-variance relationship . A random variable Y is Tweedie distributed if:

$TW_p(\mu, \sigma^2)$ if $Y\ ED(\mu, \sigma^2)$, with mean $= \mu = E(Y)$, positive dispersion parameter σ^2 and $Var(Y) = \mu\sigma^2$. The empirical model can then be written as:

$$E(Y) = f_1(\mathrm{X}_{it}) + f_n(\mathrm{Year}_t) + f_m(\mathrm{Region}_i)$$

Where the $f(.)$ stands for smooth functions; $E(Y)_{it}$ is equal to the WNV infection incidence per 100,000 in region i at time t, which we assume to be Tweedie distributed; Xit - is a vector of economic, demographic, environmental and climate variables. $Year_t$ is a function of the time intercept and $Region_i$ represents neighborhood structure of region.

Climate modeling

In order to model long term seasonal climate trends, we fit a GAM model using the following equation.

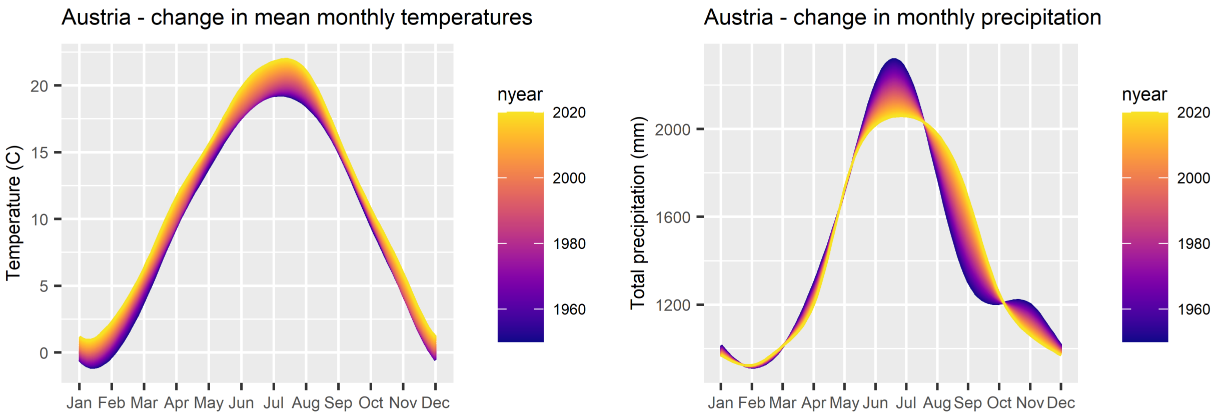
$$y = \beta_0 + f(x_1, x_2) + \varepsilon$$

where $y =$ is either the mean of the monthly regional temperatures (°C) or regional sum precipitation (mm).

B_0 is the intercept, month is represented by x_1 and x_2 is the series of years in the entire time period i.e. within-year and between year.

f is a smooth function interaction that accounts for variation in, or interaction between, the trend and seasonal features of the data.

Temperature models were fit using the Gaussian distribution and precipitation models fit using the Tweedie distribution.


Austria - change in mean monthly temperatures

Austria - change in monthly precipitation

Figure S1: Austria - seasonal climate trends (Data source: E-OBS version 22.0e).

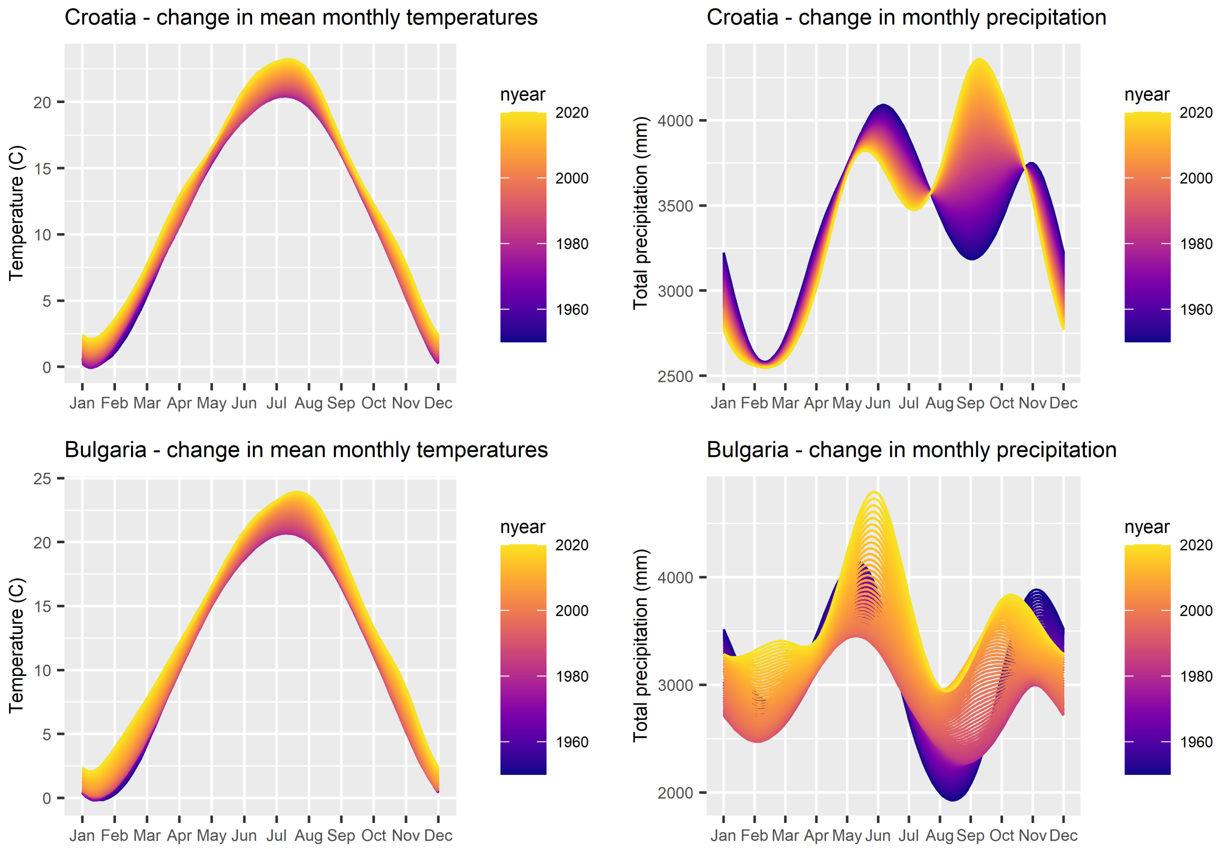


Figure S2: Bulgaria / Croatia - seasonal climate trends (Data source: E-OBS version 22.0e).

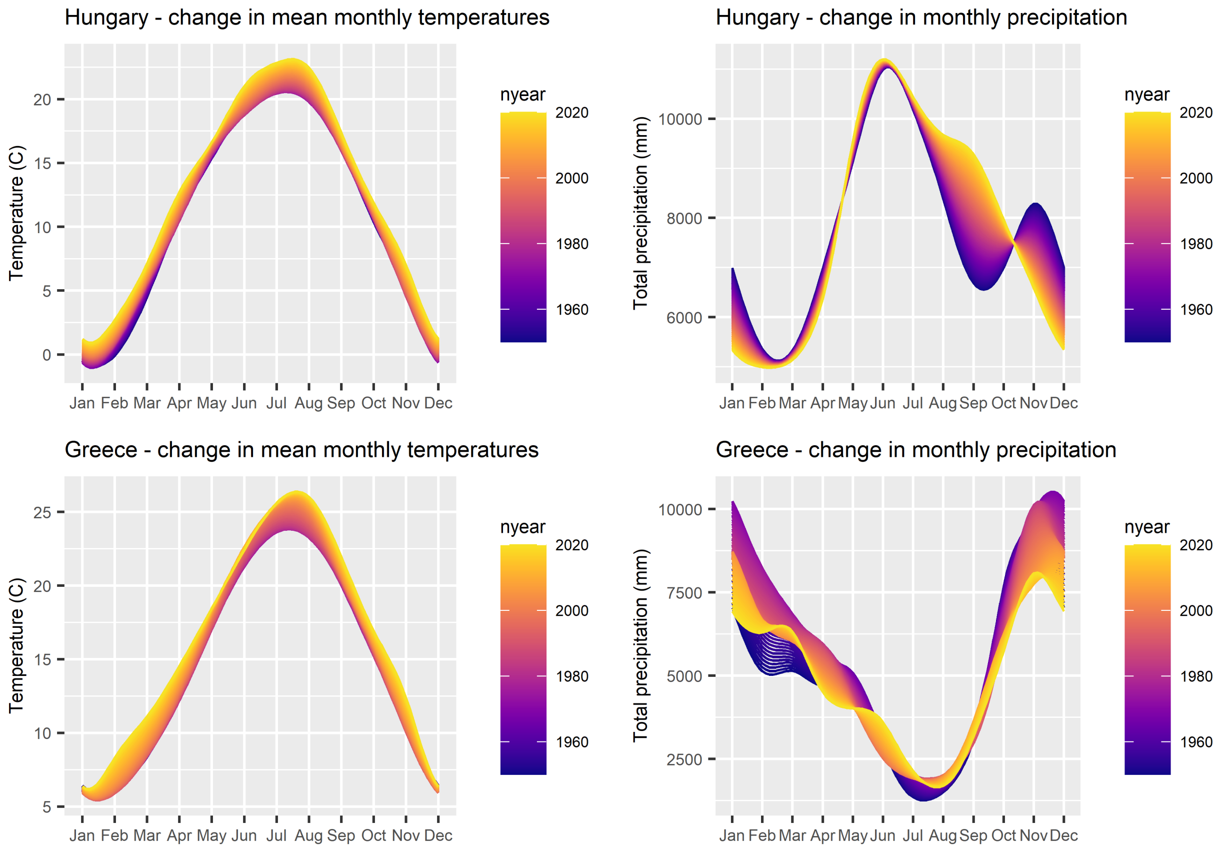


Figure S3: Greece / Hungary- seasonal climate trends (Data source: E-OBS version 22.0e).

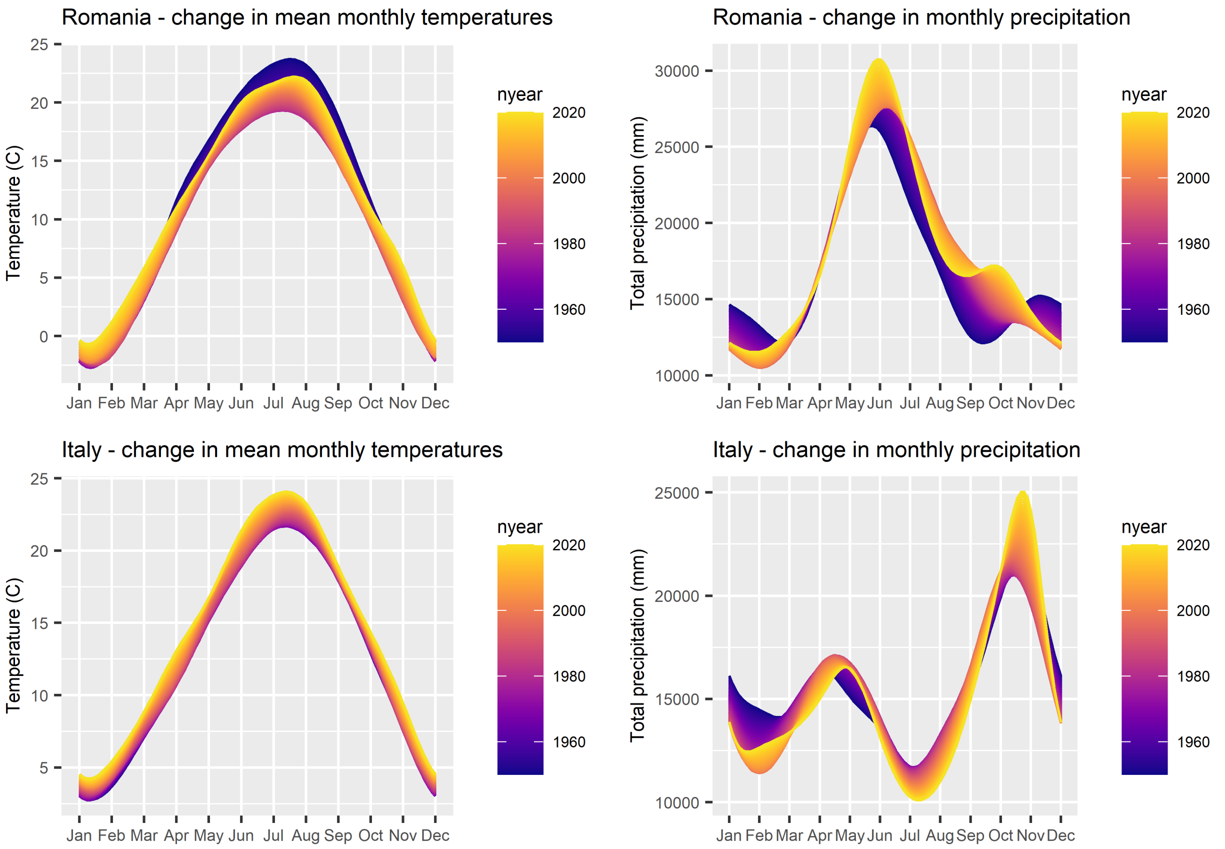


Figure S4: Romania / Italy - seasonal climate trends (Data source: E-OBS version 22.0e).

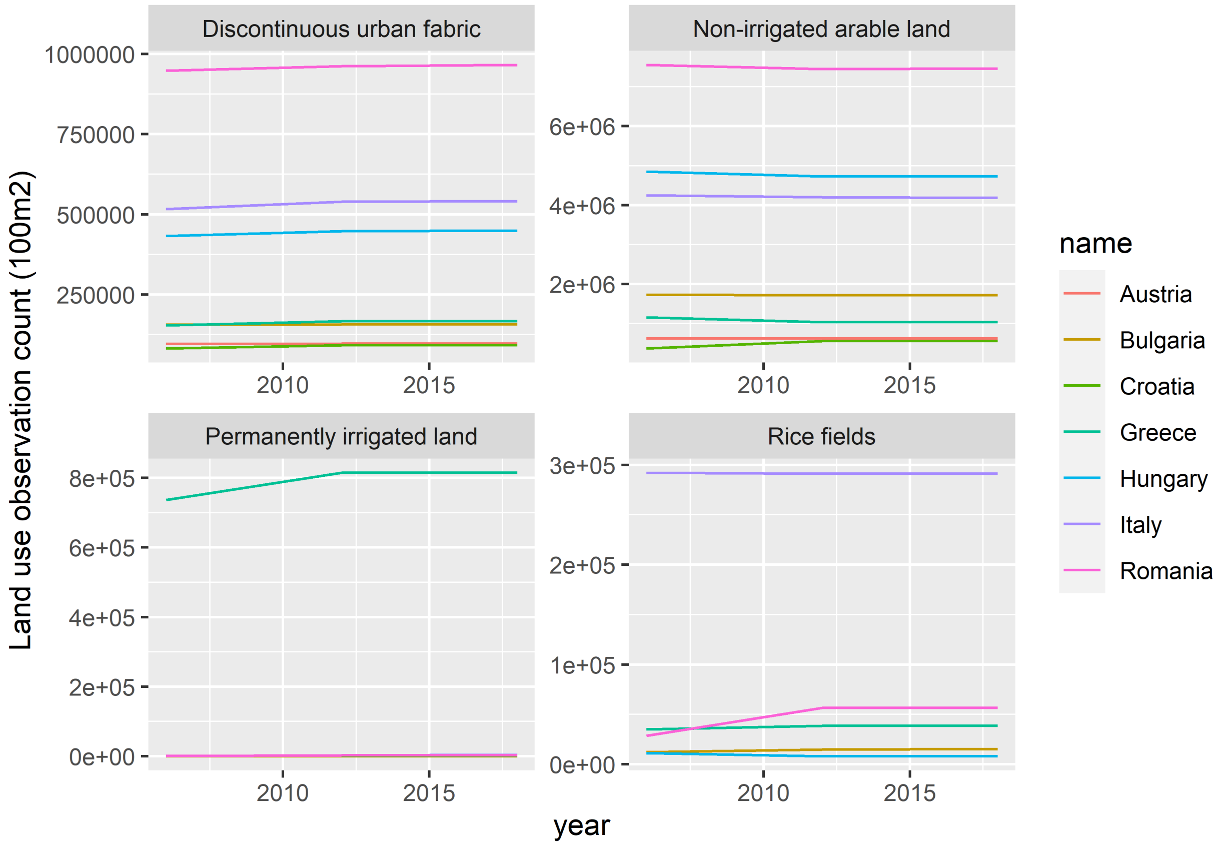


Figure S5: Land-use: 1 = Discontinuous Urban Fabric, 2-4 = Arable land break-down (Source: CORINE Land Cover)

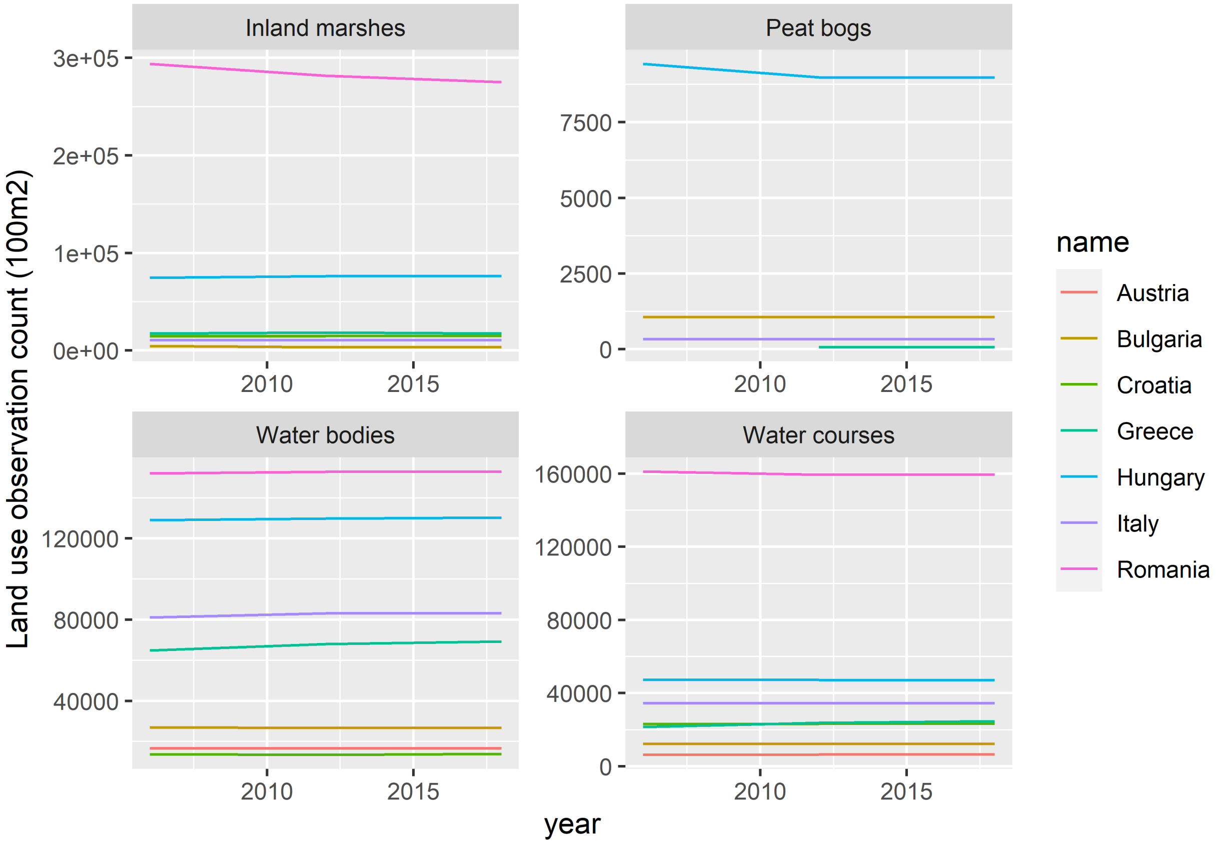


Figure S6: Land-use: Fresh water bodies break-down (Source: CORINE Land Cover)

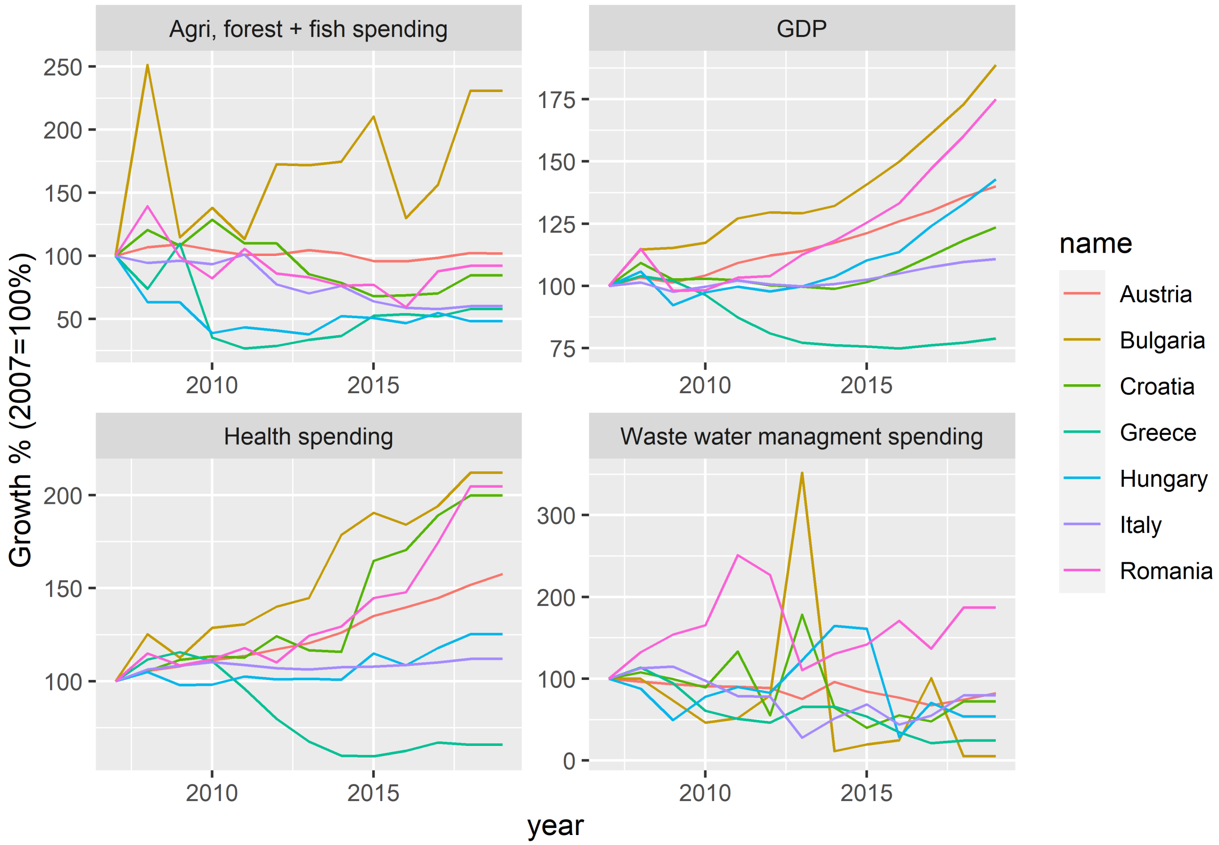


Figure S7: Government spending growth, GDP growth and unemployment 2007-2019 (2007=100%) (Source: Eurostat)

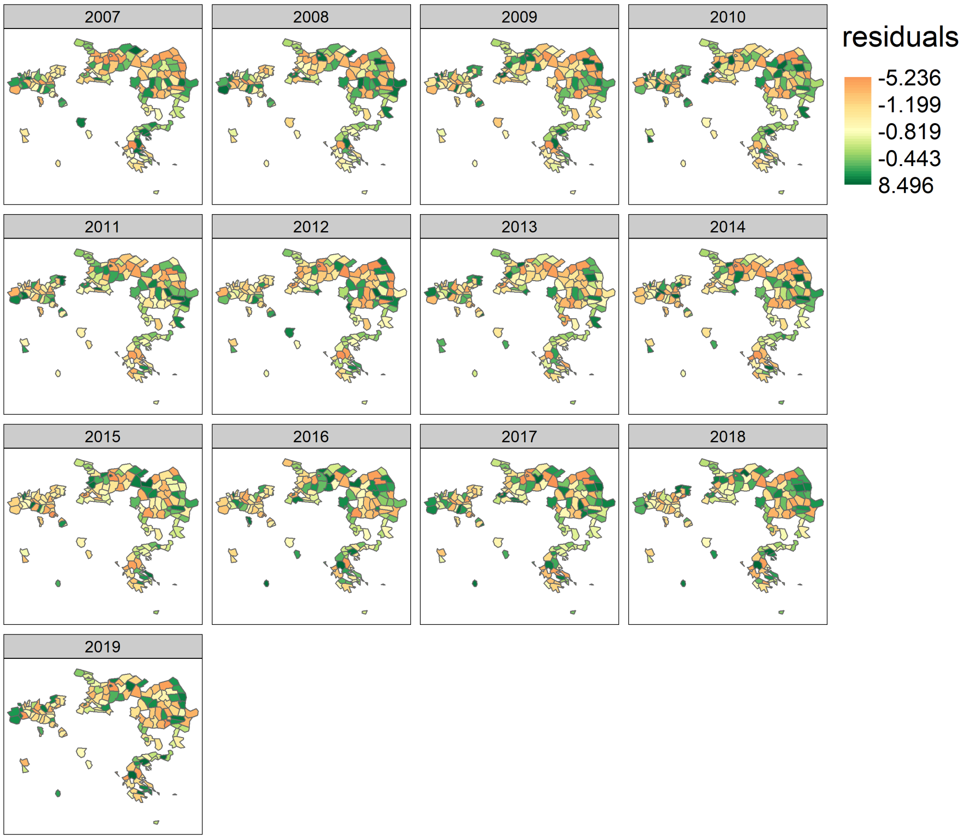


Figure S8: Variable correlation plot.

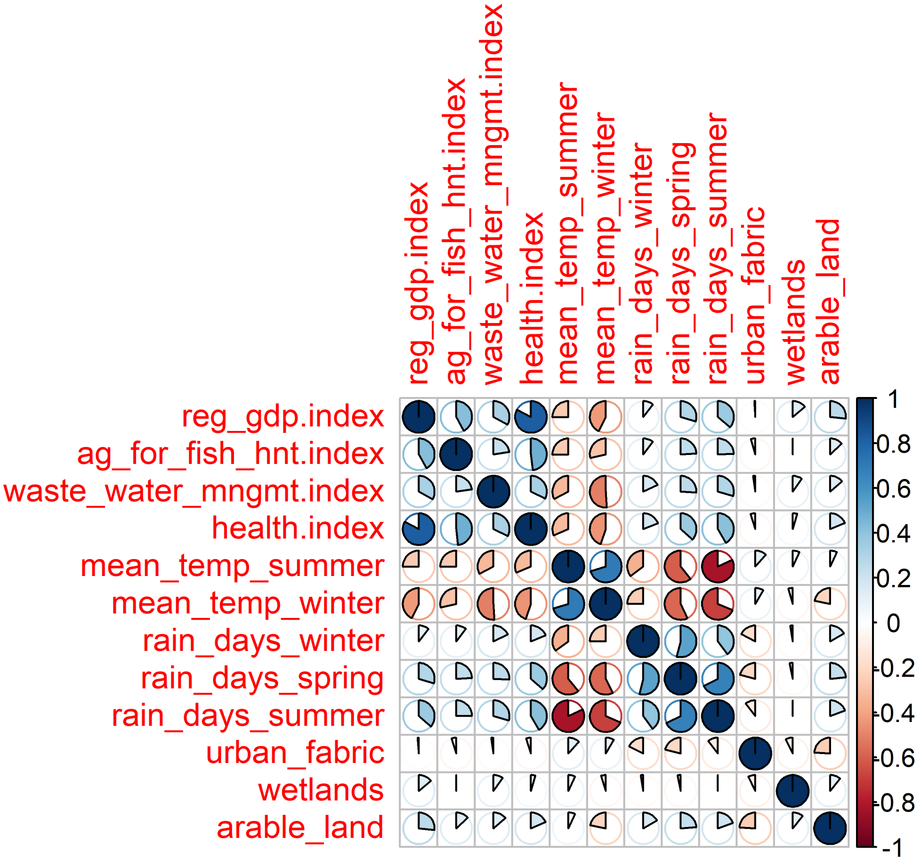


Figure S9: Spatial Residuals Tweedie model.

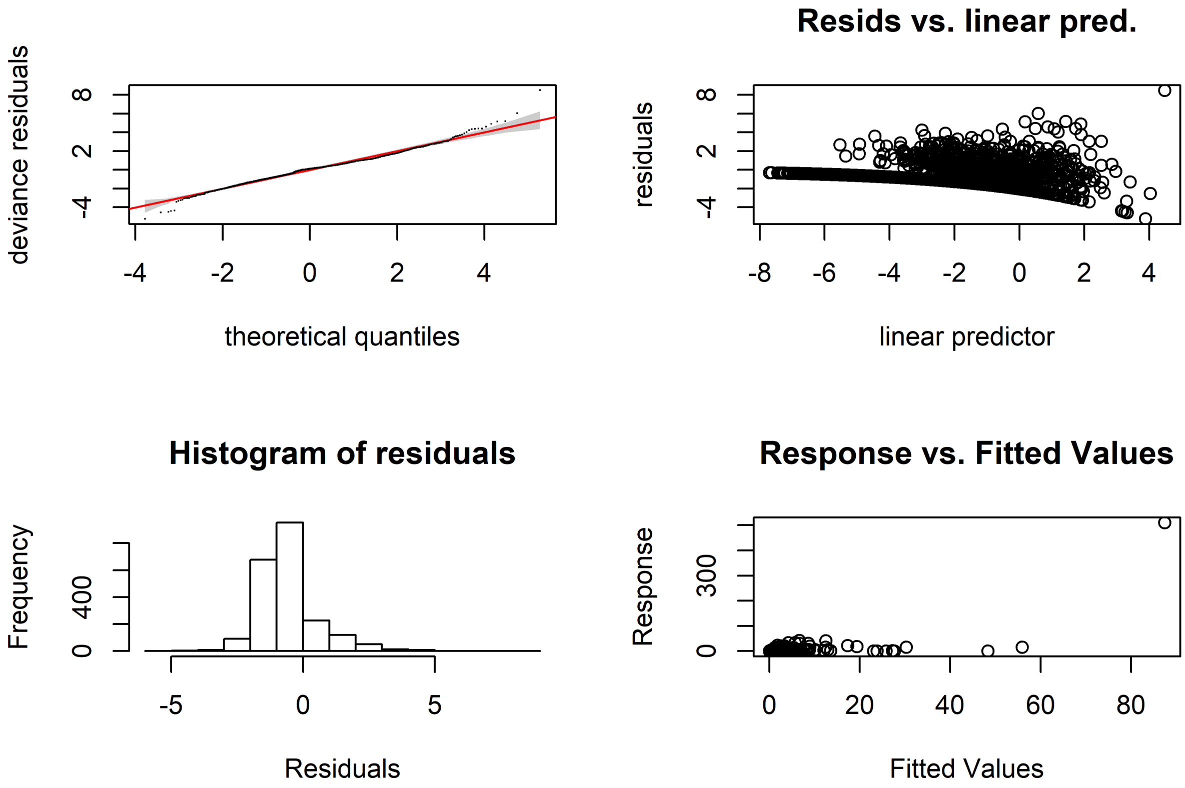


Figure S10: Diagnostics Tweedie model.

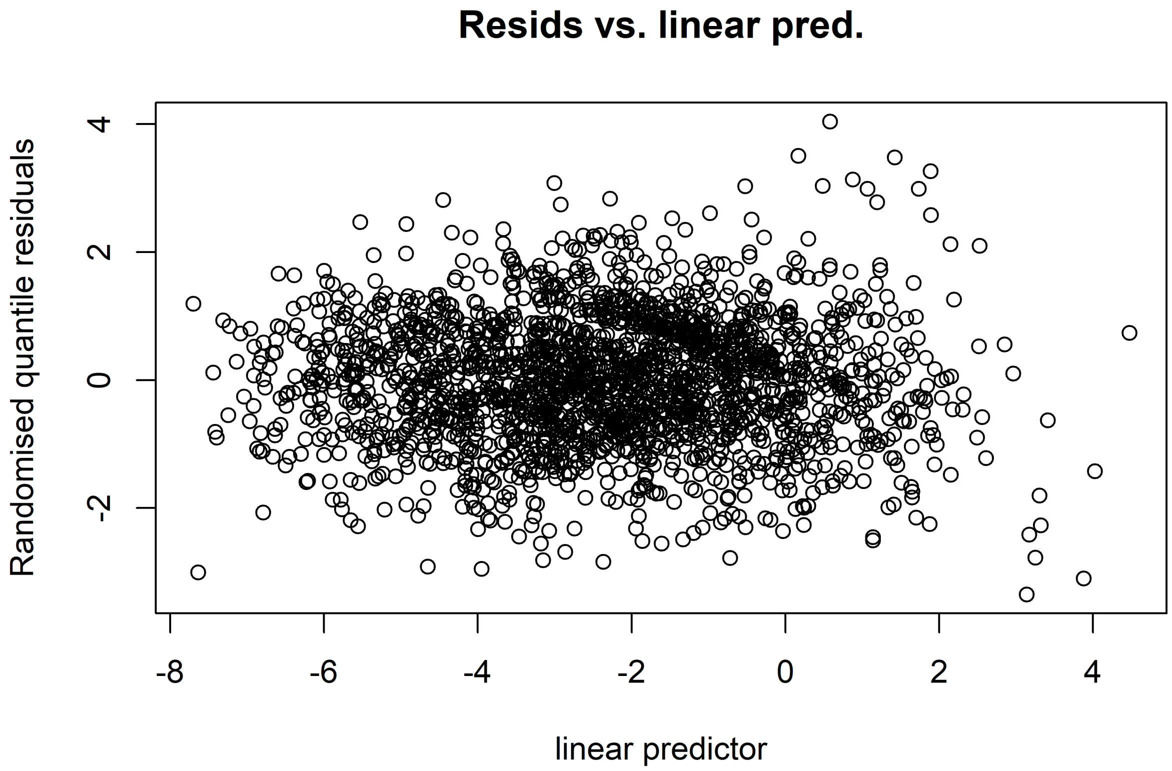


Figure S11: Diagnostics 2 Tweedie model.

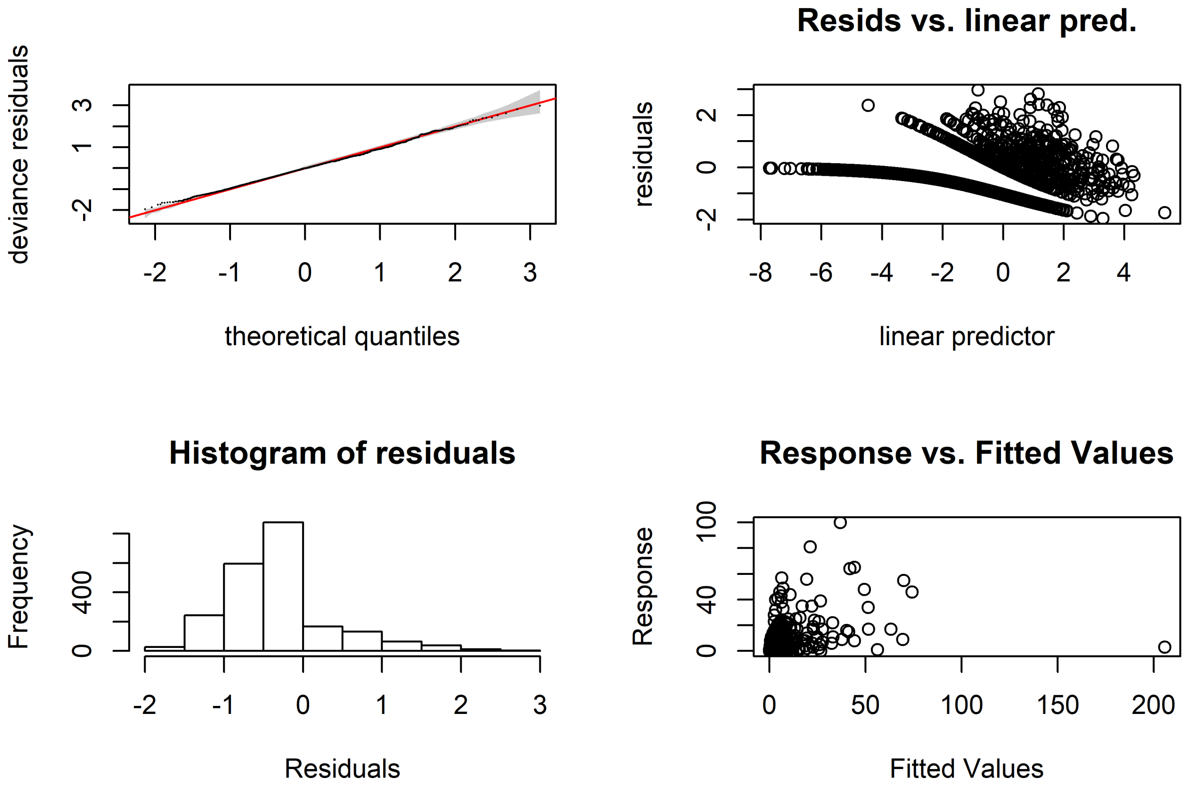


Figure S12: Diagnostics negbin model.

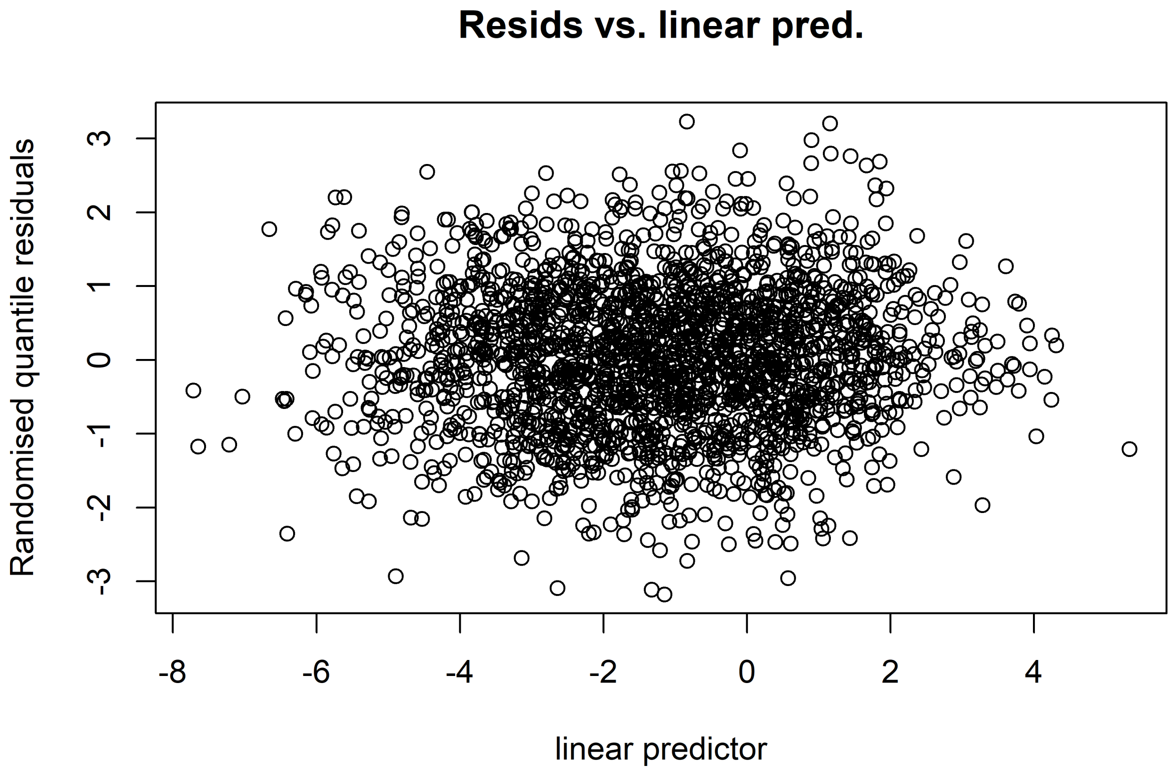


Figure S13: Diagnostics 2 negbin model.

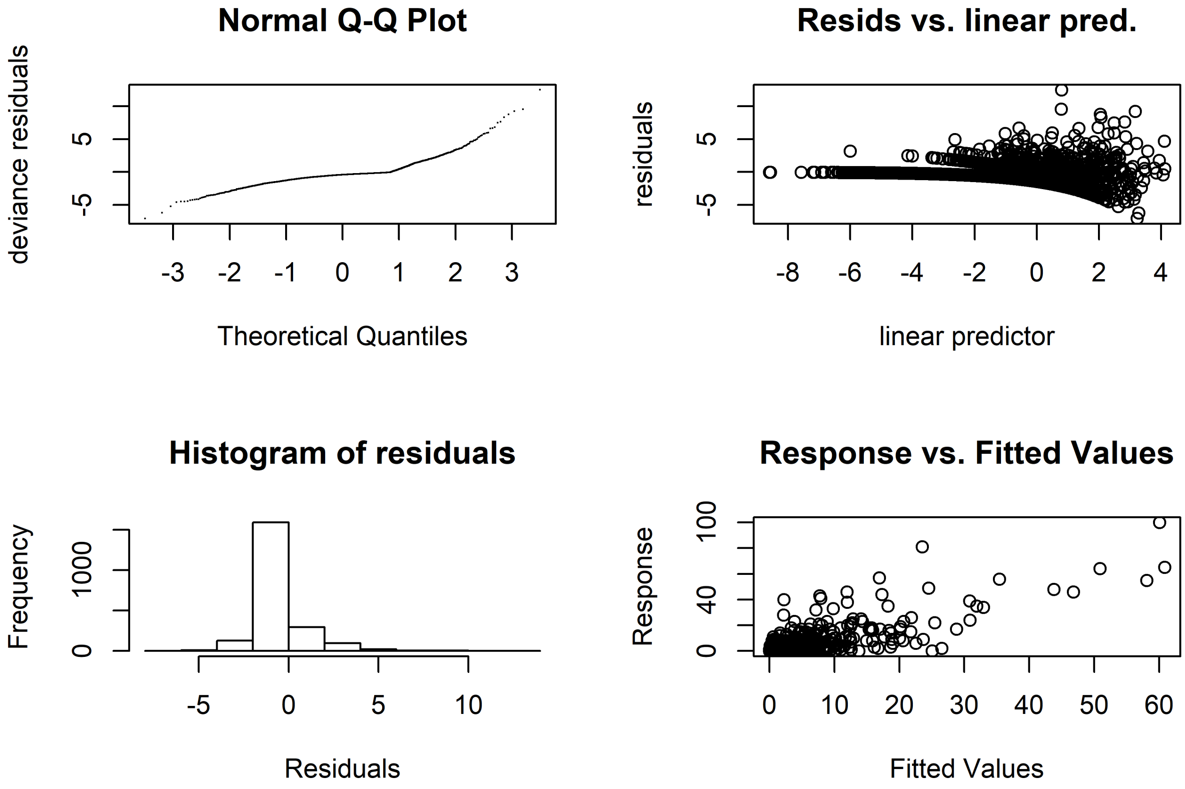


Figure S14: Diagnostics Quasipoisson model.

Table S1: WNF Cases Per Country 2006-2019

Country	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Austria	0	0	2	1	0	0	0	2	6	5	6	20	4
Bulgaria	0	0	0	0	0	2	0	1	2	1	1	15	6
Croatia	0	0	0	0	0	6	20	1	1	2	5	57	1
Cyprus	0	0	0	0	0	0	0	0	0	1	0	1	24
Czechia	0	0	0	0	0	0	1	0	0	0	0	5	1
France	0	0	0	0	0	0	0	0	1	0	2	27	1
Germany	0	0	0	0	0	0	0	0	0	0	0	1	4
Greece	0	0	0	262	100	157	85	15	0	0	48	312	228
Hungary	4	19	7	18	4	17	35	10	18	44	20	216	72
Italy	0	0	0	4	18	45	80	24	61	76	53	610	54
Portugal	0	0	0	0	0	0	0	0	1	0	0	0	0
Romania	4	2	2	57	11	15	24	23	32	93	66	279	68
Slovakia	0	0	0	0	0	0	0	0	0	0	0	0	1
Slovenia	0	0	0	0	0	0	1	0	0	0	0	4	0
Spain	0	0	0	2	0	0	0	0	0	4	0	0	0
Turkey	0	0	0	47	5	0	0	0	0	1	7	26	10

Table S2: Final model specification comparisons by distribution.v

	Tweedie model	Negbin model	Quasi Poisson model
Intercept	-2.35^{***}	-13.82^{***}	-13.86^{***}
	(0.40)	(0.36)	(0.45)
Mean temp summer (C)	1.00^{*}	1.50^{*}	2.00^{**}
	(1.00)	(1.70)	(2.00)
Mean temp winter (C)	1.94^{***}	1.96^{***}	2.00^{***}
	(1.99)	(1.99)	(2.00)
Days of rain in summer	1.00^{**}	1.00^{**}	1.00
	(1.00)	(1.00)	(1.00)
Summer surface water extent (30m2)	1.02^{***}	1.40^{*}	1.62^{***}
	(1.03)	(1.64)	(1.85)
Regional GDP index (2007=100%)	1.00	1.00	1.00
	(1.00)	(1.00)	(1.00)
Agri, forest + fish spending (2007=100%)	1.93^{***}	1.94^{***}	2.00^{***}
	(1.99)	(1.99)	(2.00)
Waste water managment spending (2007=100%)	1.10^{***}	1.45^{***}	1.75^{***}
	(1.19)	(1.69)	(1.93)
Continuous urban fabric %	1.00	1.00	1.00
	(1.00)	(1.00)	(1.00)
Discontinuous urban fabric %	1.00	1.00	1.00
	(1.00)	(1.00)	(1.00)
Wetlands %	1.00	1.00	1.01
	(1.00)	(1.00)	(1.01)
Arable land %	1.74^{**}	1.83^{***}	1.00^{**}
	(1.84)	(1.90)	(1.00)
Year	11.56^{***}	11.55^{***}	11.62^{***}
	(12.00)	(12.00)	(12.00)
Spatial lag	76.19^{***}	83.79^{***}	117.20^{***}
	(106.52)	(115.75)	(141.05)
AIC	3907.56	4659.28	-
BIC	4538.85	5335.48	-
Log Likelihood	-1842.57	-2210.53	-
Deviance	3520.85	1207.27	4607.49
Deviance explained	0.65	0.64	0.69
Dispersion	2.73	1.00	3.22
R^2	0.25	-0.32	0.60
GCV score	1871.90	2367.22	2.45
Num. obs.	2158	2158	2158
Num. smooth terms	13	13	13

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

References

A. Aswi, S. M. Cramb, P. Moraga, and K. Mengersen. Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review. *Epidemiology and infection*, 147:1–14, 2018. ISSN 1469-4409 0950-2688. doi: 10.1017/S0950268818002807. URL https://pubmed.ncbi.nlm.nih.gov/30369335https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6518570/.

R. S. Bivand, E. Pebesma, and V. Gomez-Rubio. *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013. URL https://asdar-book.org/.

R. C. Cornes, G. van der Schrier, E. J. M. van den Besselaar, and P. D. Jones. An ensemble version of the e-obs temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, 123(17):9391–9409, 2018. ISSN 2169-897X. doi: https://doi.org/10.1029/2017JD028200. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2017JD028200.

EU. Copernicus land monitoring service 2018, 2018. URL https://www.eea.europa.eu/data-and-maps/data/copernicus-land-monitoring-service-corine.

Eurostat. Agriculture, forestry and fishery statistics. Report, Eurostat, 2019. URL https://ec.europa.eu/eurostat/web/products-statistical-books/-/KS-FK-19-001.

Eurostat. Nuts - nomenclature of territorial units for statistics, 2020. URL https://ec.europa.eu/eurostat/web/nuts/background.

C. Kurz. Tweedie distributions for fitting semicontinuous health care utilization cost data. *BMC Medical Research Methodology*, 17(1), 12 2017. doi: 10.1186/s12874-017-0445-y. URL http://dx.doi.org/10.1186/s12874-017-0445-y.

L. Lahti, J. Huovari, M. Kainu, and P. Biecek. eurostat r package, 2017. URL https://journal.r-project.org/archive/2017/RJ-2017-019/index.html. Version 3.7.5.

E. Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1): 439–446, 2018. doi: 10.32614/RJ-2018-009. URL https://doi.org/10.32614/RJ-2018-009.

J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward. High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633):418–422, 2016. ISSN 1476-4687. doi: 10.1038/nature20584. URL https://doi.org/10.1038/nature20584.

R. J. H. . J. van Etten. *raster: Geographic analysis and modeling with raster data*, 2012. URL http://CRAN.R-project.org/package=raster. R package version 2.0-12.

S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, 2011.

S. N. Wood. *Generalized additive models: an introduction with R*. CRC press, 2017. ISBN 1498728340.

A. Zeileis and G. Grothendieck. zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27, 2005. doi: 10.18637/jss.v014.i06.