A unified framework to model the potential and realized distributions of invasive species within the invaded range by Hattab *et al*.

Appendix S1: Description of the predictor variables used in this study

We used a very comprehensive set of 22 predictor variables available as vector or fine-grained (0.5-m to 1.5-m) raster maps across the entire study area, including: topographic variables; soil properties variables; resource availability variables; biotic variables; and anthropogenic variables (see Table 1 in the main manuscript for the full list of variables). These variables were used as predictor in our species distribution models (SDMs) to model the potential and realized distributions of invasive species, being either a virtual or a real invasive species. Sixteen raster maps were derived at 0.5-m resolution from airborne light detection and ranging (LiDAR) data (see Appendix S2 for details about LiDAR data acquisition and processing and Appendix S3 for maps). The NDVI map was obtained from a remotely sensed SPOT6 satellite image acquired in June 2014 at 1.5-m resolution (http://ids.equipex-geosud.fr/). Moreover, a 25-m resolution soil pH (H2O) map was generated using 166 field-based measurements of pH in the topsoil by applying the generic framework for spatial prediction of soil variables based on regression-kriging (Hengl *et al.*, 2004) (See Appendix S2 for details). A vector map with information on stand age (an interval variable), stand type (a factor variable of the dominant tree species in each stand) and silvicultural practices (a binary variable: managed *vs*. unmanaged) as well as a vector map with information on soil types were provided by the French forests national office (ONF).

Appendix S2: Light detection and ranging (LiDAR) data acquisition and processing

Airborne light detection and ranging (LiDAR) data was used to derive various landscape and canopy structure maps across the entire forest of Compiègne in northern France (see Fig. 2a in the main manuscript). The AERODATA Company (http://www.aerodata-france.com/) installed a Riegl LMS-680i laser scanner (http://www.riegl.com/) on a plane and performed flights in February 2014 to get an average density of 12 points.m^{-2}. The AERODATA Company also performed the post-processing of the data to calibrate and merge laser bands, to create a raw point cloud and to classify each point as "ground", "vegetation", "water" or "building". Data classification was performed using an automatic ground search algorithm. This routine starts a search procedure to find the lowest last-return points within grid cells of 100 m^2 each (cf. ground points), which are then connected by a triangulated irregular network (TIN). This initial gross TIN is iteratively refined by adding likely ground points that fulfil threshold conditions in terms of distance and angle towards the TIN model. The iterative point addition process is stopped once no more points fulfil the threshold conditions. We then interpolated the points classified as ground returns and the unclassified points to create respectively a 0.5-m digital terrain model (DTM) and digital elevation model (DEM) both comprising 45360 × 45360 cells: about 2 giga-cells.

Using the DTM and the DEM, we created a set of 17 environmental variables belonging to five types of variables: topographical variables; resource availability variables; biotic variables; anthropogenic variables; and soil properties variables. Topographical variables were calculated from the DTM, including: elevation; slope; aspect; and tangential curvature (the curvatures in the direction of the contour tangent). To derive this set of four topographical variables, we used topographical modelling features under GRASS GIS 7 (Neteler *et al.*, 2012). Thereafter we transformed the aspect into two derived variables: eastness (values close to 1 represent an eastward aspect, while values close to –1 represent a westward aspect) and northness (values close to 1 represent a northward aspect, while values close to –1 represent a southward aspect).

In order to quantify topographic control on hydrological processes, we also derived from the DTM: the topographic wetness index and the flow accumulation index. To derive these hydrographic variables, we used hydrographical modelling tools under GRASS GIS 7 (Neteler *et al.*, 2012). Besides, we used the DEM to quantify the incoming solar radiation: the potential annual global radiation and the average hours of sun per day. To do so, we used the algorithm described in (Šúri & Hofierka, 2004), as implemented in the "r.sun" command running under GRASS GIS 7 (Neteler *et al.*, 2012). It computes clear-sky solar radiation at a daily time step using the model of the European Solar Radiation Atlas (http://www.soda-is.com/esra: (Rigollier *et al.*, 2000)), which is based on the Linke turbidity coefficient. This command also includes routines to account for hillshading effects caused by variations in solar angle, ground slope and aspect as well as shadowing effects of adjacent topographic features. Note that the Linke turbidity coefficient represents the transparency of the cloudless atmosphere by accounting for aerosols and water vapour in the atmosphere.

A suite of forest canopy metrics was derived from LiDAR data to quantify different aspects of forest 3-D structure including the canopy height profiles and the canopy density. To achieve this, we first subtracted elevation values of the LiDAR point-returns classified as "vegetation" from the 0.5-m DTM. This provides a measure of vegetation height. The vegetation structural characteristics were then summarized by six distributional statistics calculated within a 0.5-m spatial unit from the vegetation height returns: minimum; the 5th percentile; mean; the 95th percentile; maximum; and standard deviation. In addition to that, we divided the total number of points classified as "vegetation" by the total number of points within each 0.5-m spatial unit, to derive an index of canopy density ranging between 0 and 1.

For soil data, a 25-m resolution map of soil pH (H_2O) map was generated using 166 field-based measurements of pH in the topsoil. These field-based measurements were collected following an environmental systematic survey (see Figs. 1 and 2 as well as the "Application using a real invasive species" section in the main manuscript. At each plot, we collected 10 soil samples of the topsoil (15 cm depth after removal of soil litter). Finally, soil pH (H_2O) of the composite soil samples was measured in a water suspension following the norm ISO 10390 (AFNOR, 1996).To obtain the soil pH (H_2O) map, we applied a generic framework for spatial prediction of soil variables based on regression-kriging (Hengl *et al.*, 2004). Regression-kriging is a spatial interpolation technique that combines a regression of the dependent variable on auxiliary variables with simple kriging of the regression residuals (Hengl *et al.*, 2004). This method has the advantage that it explicitly separates trend estimation from residual interpolation and allows the separate interpretation of the two interpolated components and explicitly incorporates spatial dependence into predictions (Hengl *et al.*, 2007). We used five variables known to influence soil pH (Tsui *et al.*, 2004; Williard *et al.*, 2005) as predictors in the trend estimation component, including: soil type; elevation; slope; a litter quality index score (this variable was derived from the most dominant tree species in a given stand type using the litter quality index scores available in (Verheyen *et al.*, 2012)); and bedrock geology (this map was based on information available at the French geological survey (BRGM)). We used a random forest algorithm in the trend estimation components since it allows capturing high non-linear and complex interactions between predictor variables (Breiman, 2001). The regression tree modelling was supplemented with the use of variograms to assess the level of spatial dependence in the residuals. Regression tree residuals were then krigged and added back to the random forest estimate. Finally, the predictive accuracy of the regression-kriging procedure was evaluated by a 10-fold cross validation. The regression-kriging predictions resulted in high prediction accuracy as indicated by a root mean square of error of 0.48 pH unit and a Pearson's correlation coefficients between the observed and the predicted pH values of 0.94.

References

AFNOR (1996) *Qualité des sols*, Association Française de Normalisation. Paris.

Breiman, L. (2001) Random forests. *Machine learning*, **45**, 5–32.

Hengl, T., Heuvelink, G.B. & Stein, A. (2004) A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, **120**, 75–93.

Hengl, T., Heuvelink, G.B.M. & Rossiter, D.G. (2007) About regression-kriging: From equations to case studies. *Computers & Geosciences*, **33**, 1301–1315.

Neteler, M., Bowman, M.H., Landa, M. & Metz, M. (2012) GRASS GIS: A multi-purpose open source GIS. *Environmental Modelling & Software*, **31**, 124–130.

Rigollier, C., Bauer, O. & Wald, L. (2000) On the clear sky model of the ESRA—European Solar Radiation Atlas—with respect to the Heliosat method. *Solar energy*, **68**, 33–48.

Šúri, M. & Hofierka, J. (2004) A new GIS-based solar radiation model and its application to photovoltaic assessments. *Transactions in GIS*, **8**, 175–190.

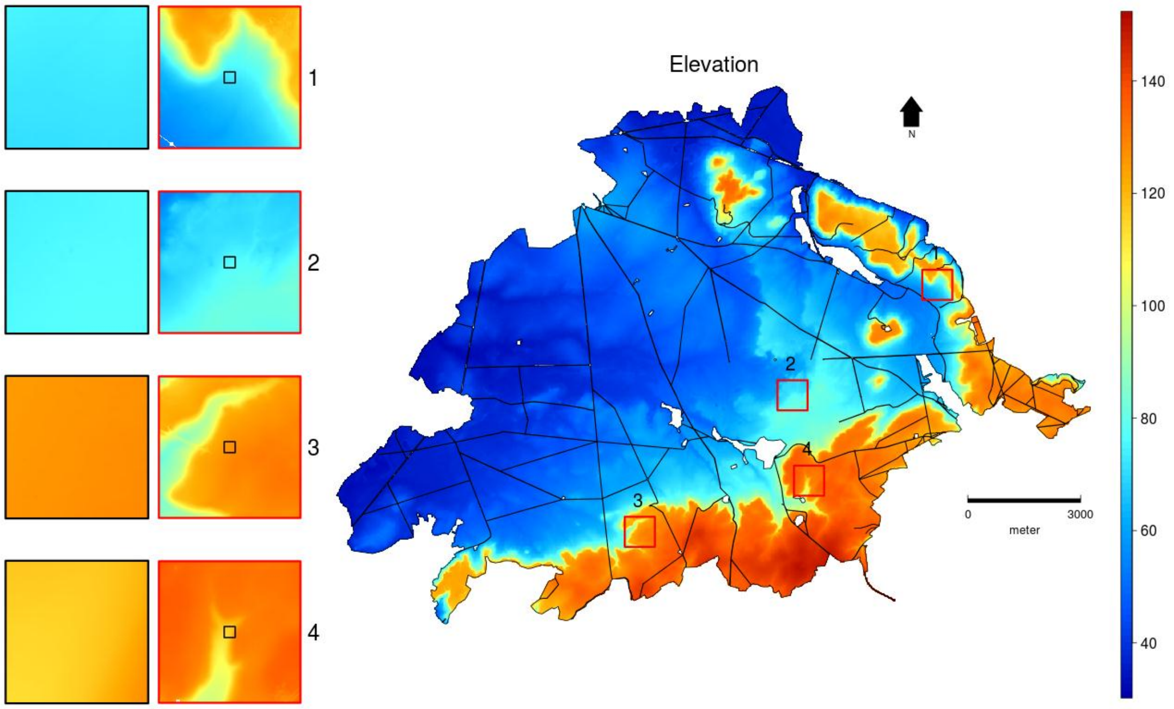
Tsui, C.-C., Chen, Z.-S. & Hsieh, C.-F. (2004) Relationships between soil properties and slope position in a lowland rain forest of southern Taiwan. *Geoderma*, **123**, 131–142.

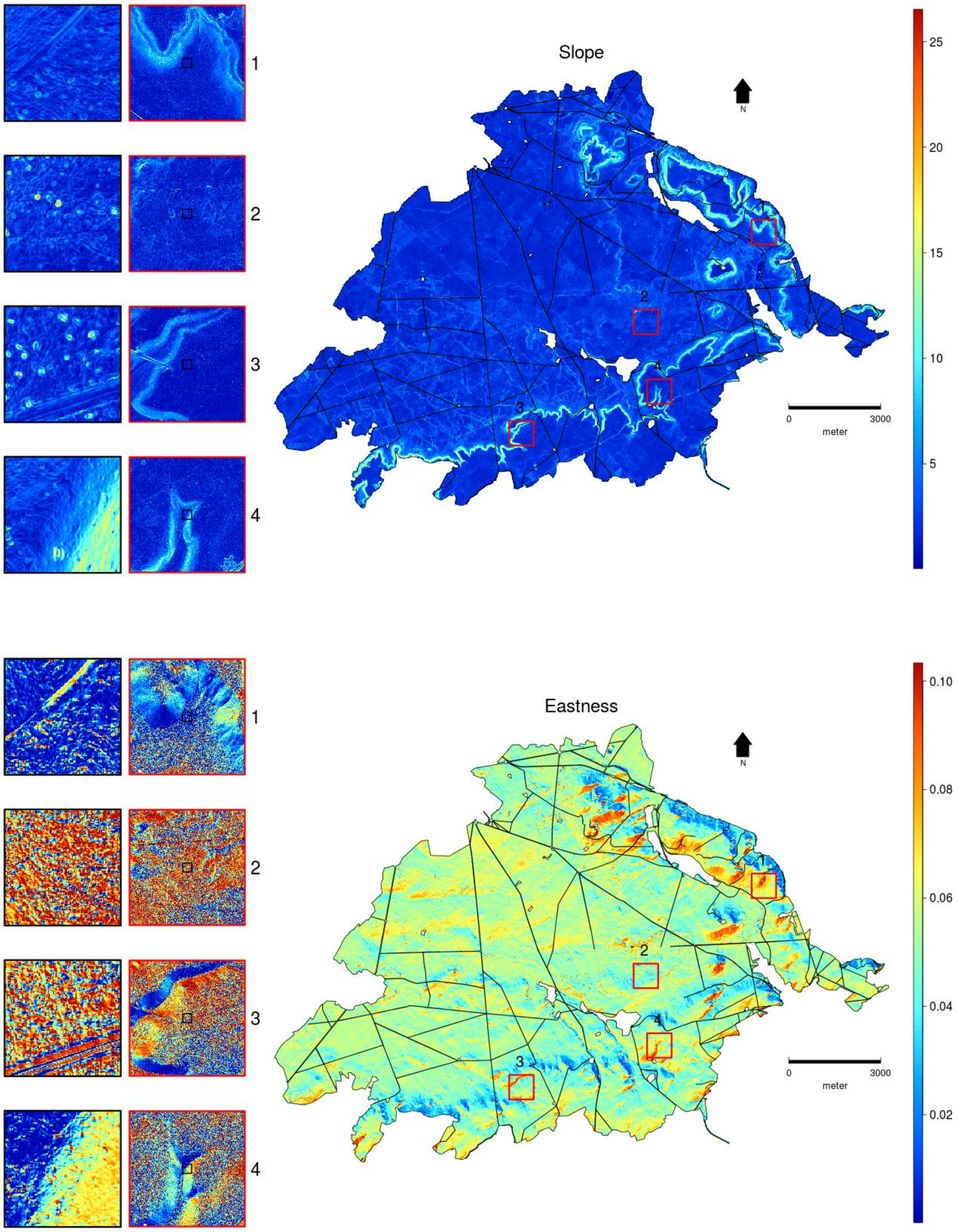
Verheyen, K., Baeten, L., De Frenne, P., Bernhardt-Römermann, M., Brunet, J., Cornelis, J., Decocq, G., Dierschke, H., Eriksson, O., Hedl, R. & others (2012) Driving factors behind the eutrophication signal in understorey plant communities of deciduous temperate forests. *Journal of Ecology*, **100**, 352–365.

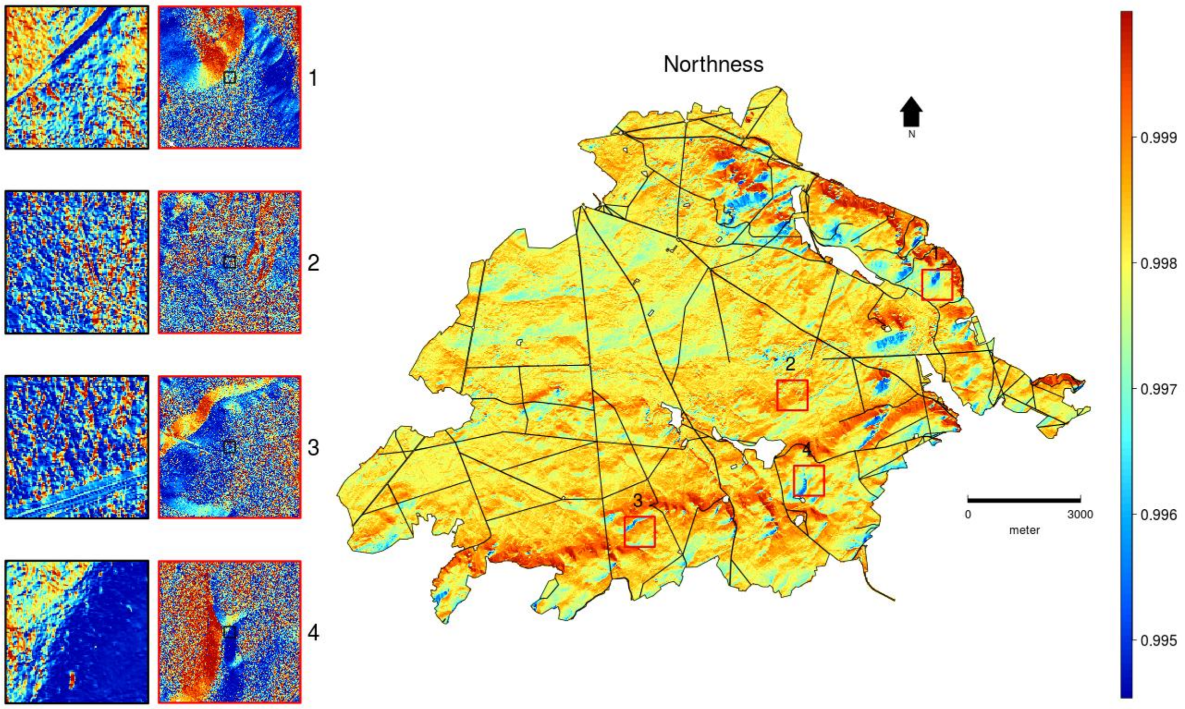
Williard, K.W., Dewalle, D.R. & Edwards, P.J. (2005) Influence of bedrock geology and tree species composition on stream nitrate concentrations in mid-Appalachian forested watersheds. *Water, Air, and Soil Pollution*, **160**, 55–76.

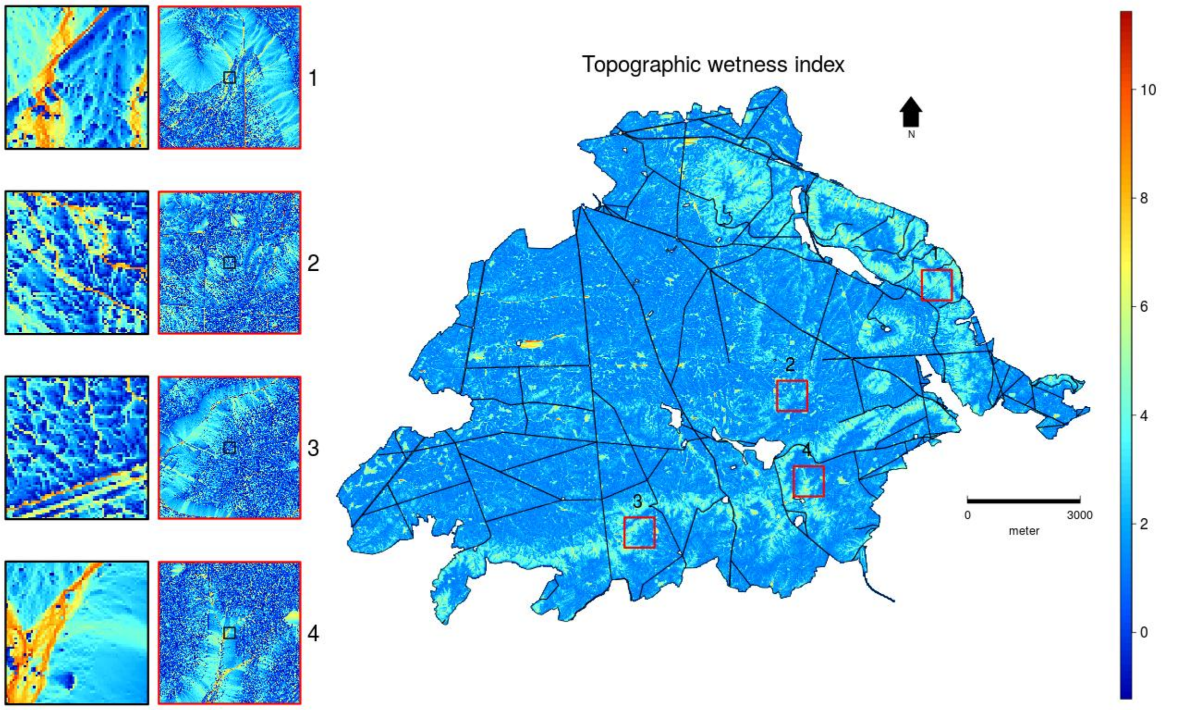
Appendix S3: Maps of predictor variables derived from airborne light detection and ranging (LiDAR) data

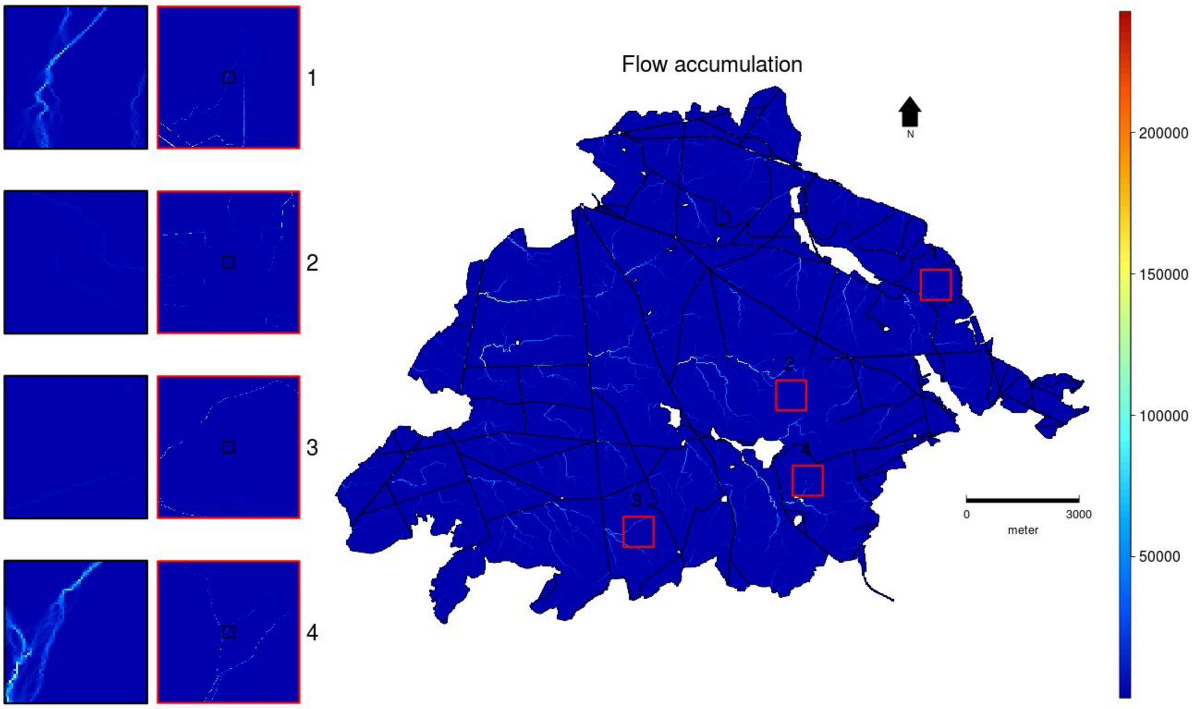
NB: All maps have a spatial resolution of 25 m across the forest of Compiègne in northern France while the sets of two cascading zooming windows have a spatial resolution of 50 cm.



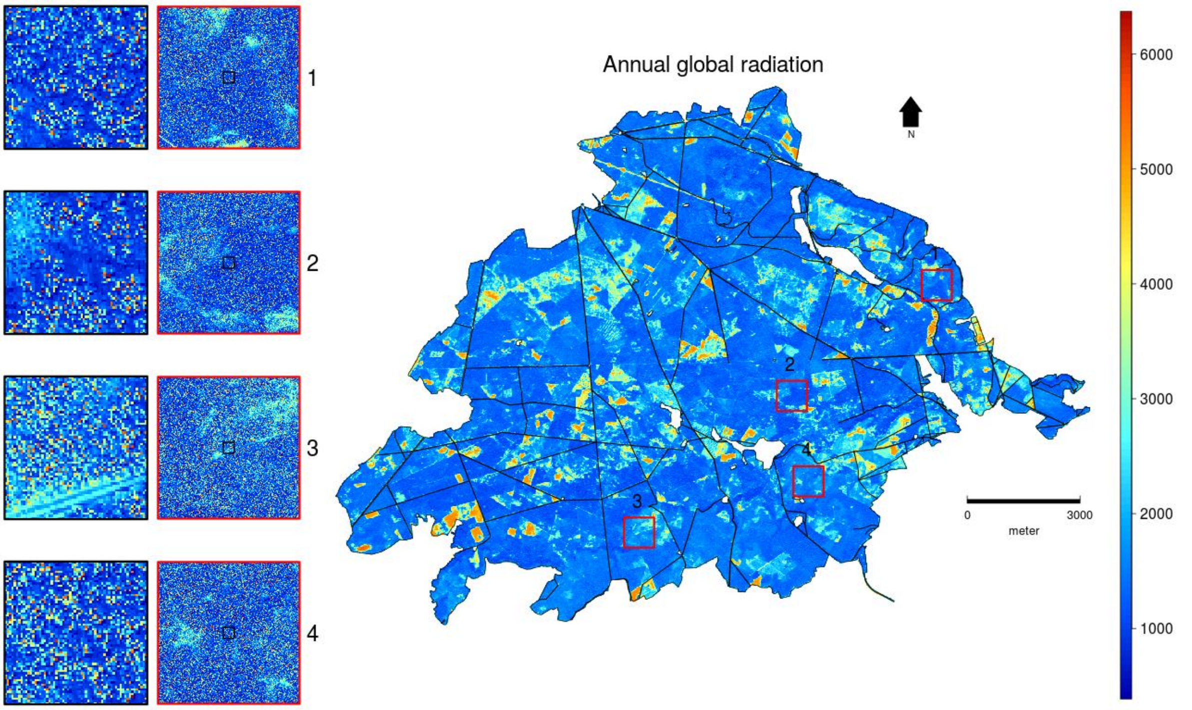

Slope
1
2
3
4
N
0
3000
meter
25
20
15
10
5
Eastness
1
2
3
4
N
0
3000
meter
0.10
0.08
0.06
0.04
0.02


1
2
3
4
Northness
N
1
2
4
3
0
3000
meter
0.999
0.998
0.997
0.996
0.995

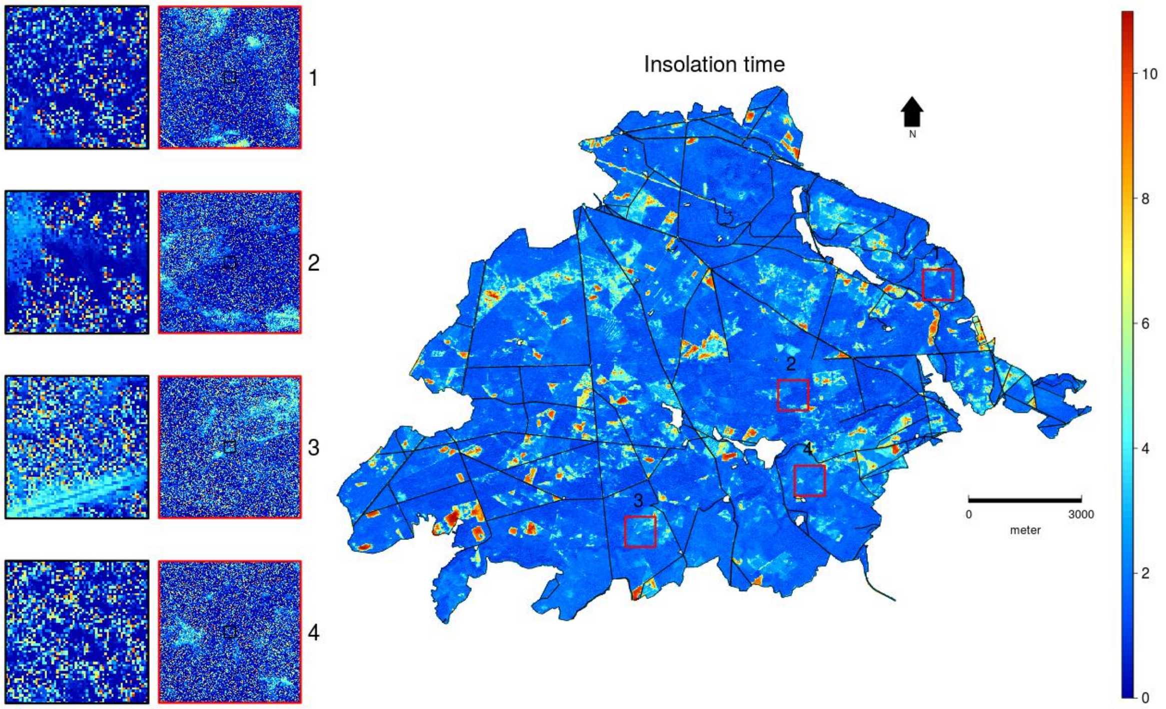

1
2
3
4
Topographic wetness index
N
1
2
4
3
0
3000
meter
10
8
6
4
2
0

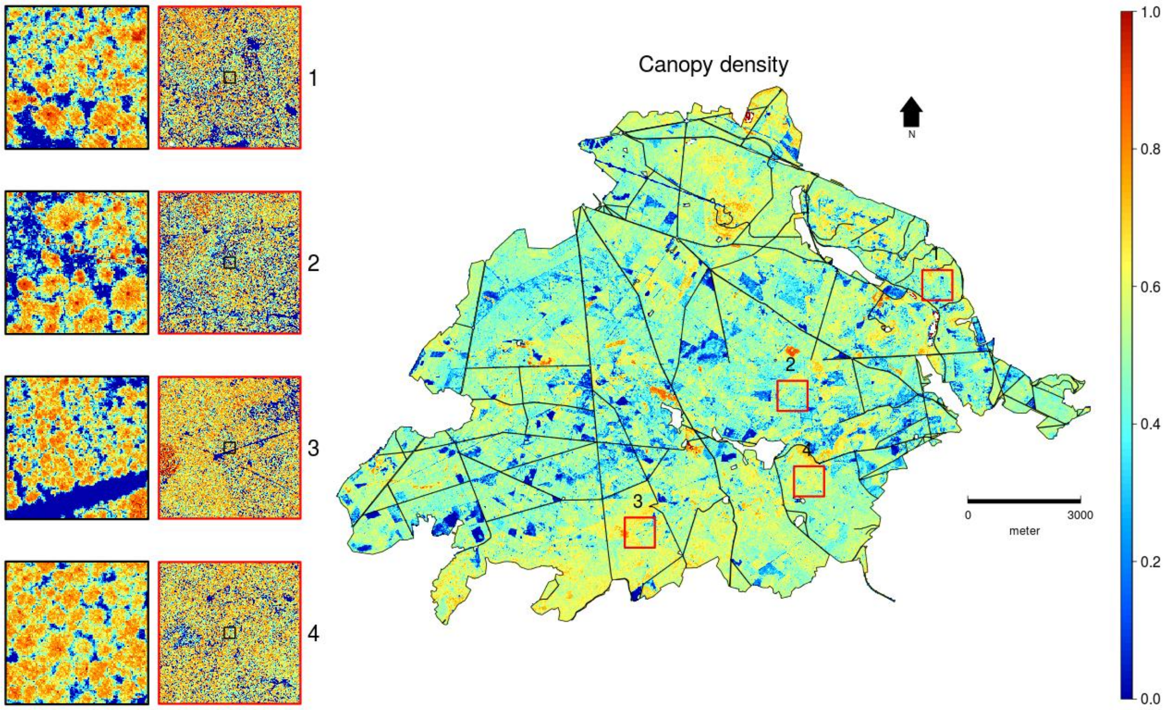


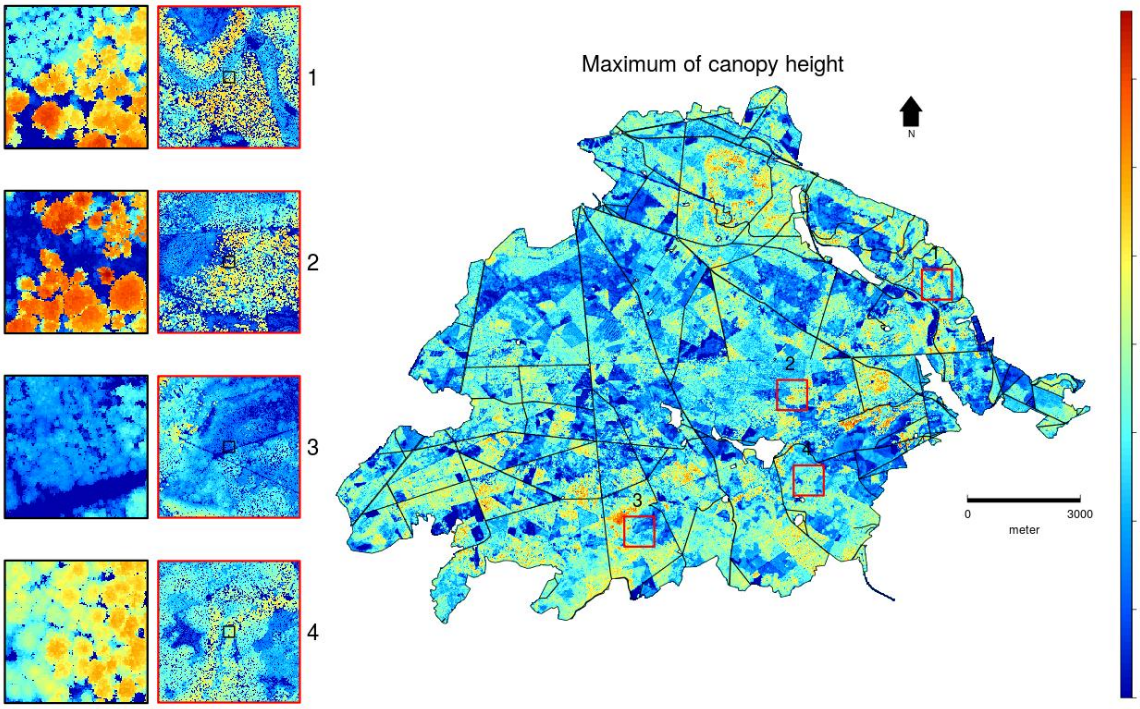
Flow accumulation



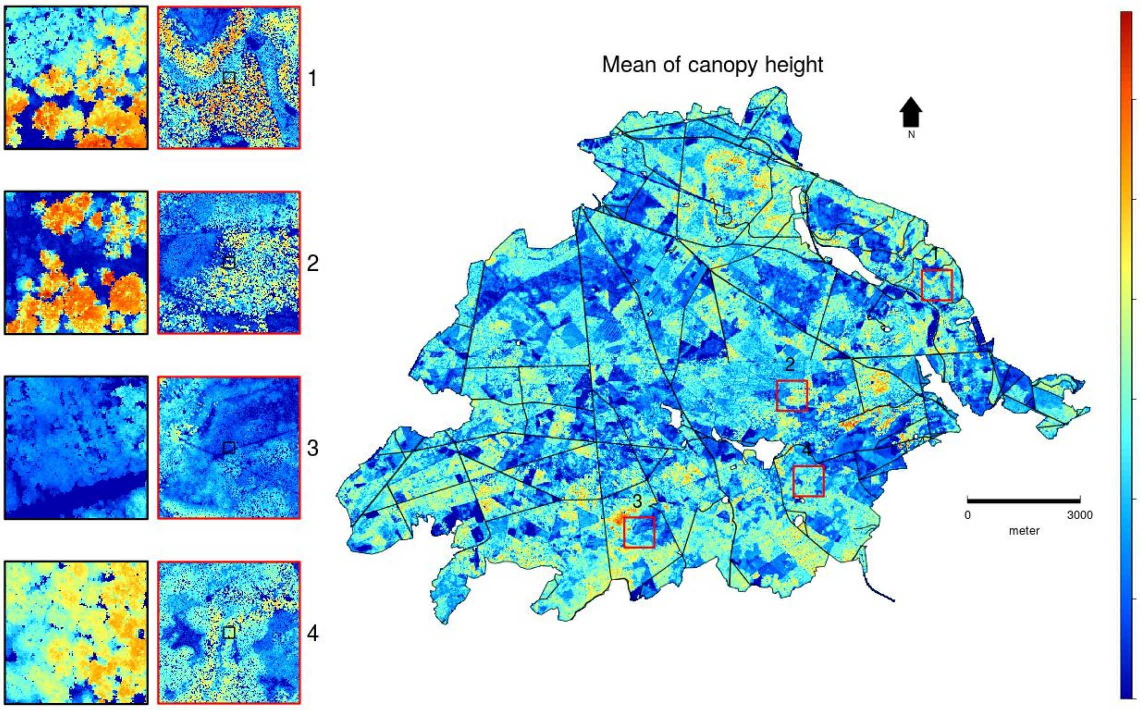
Annual global radiation


Insolation time
1
2
3
4
N
0
3000
meter
10
8
6
4
2
0

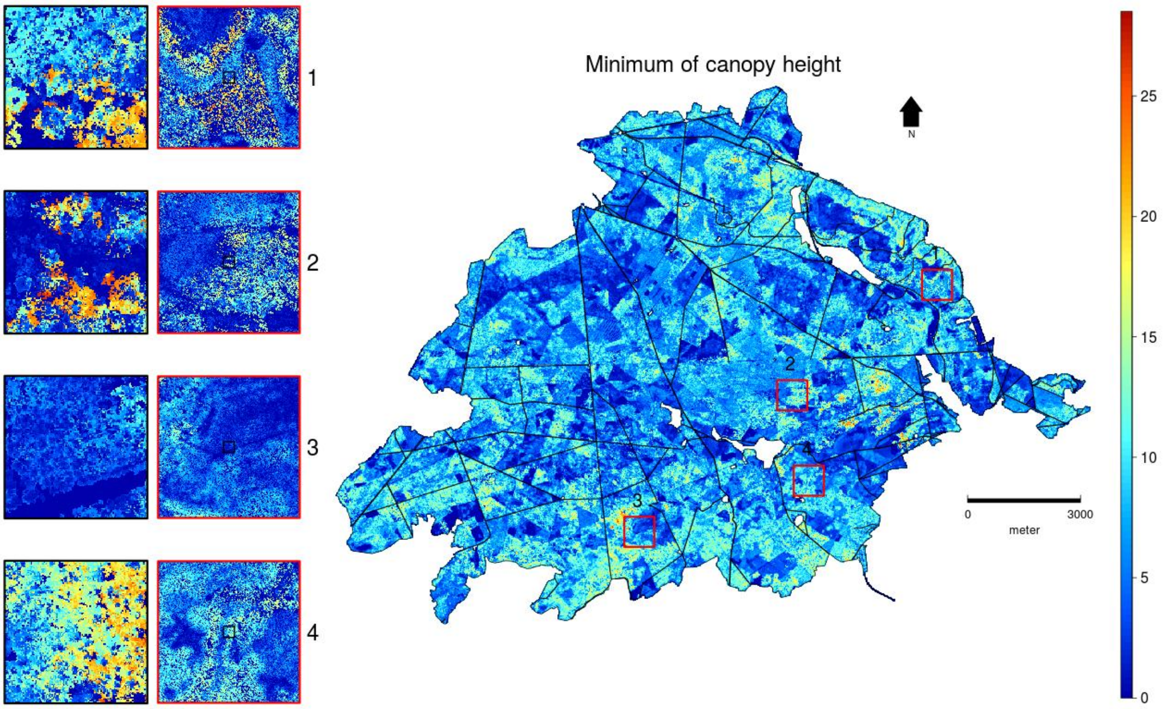

Canopy density
1
2
3
4
N
0
3000
meter
1.0
0.8
0.6
0.4
0.2
0.0

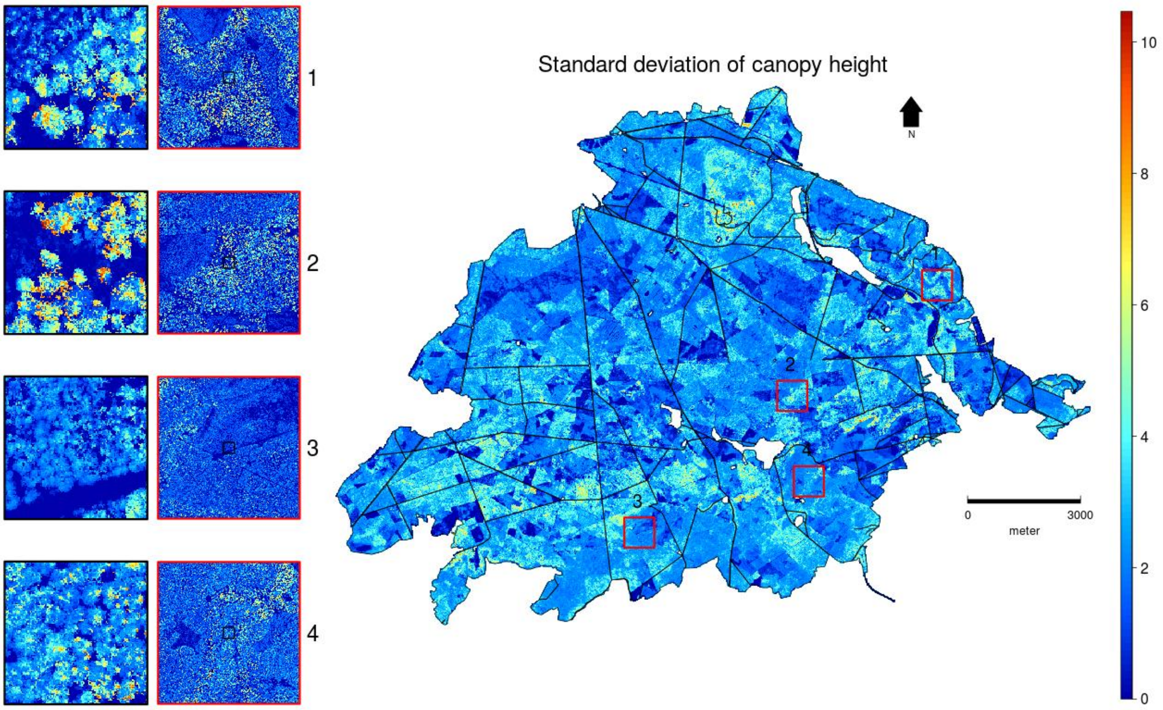


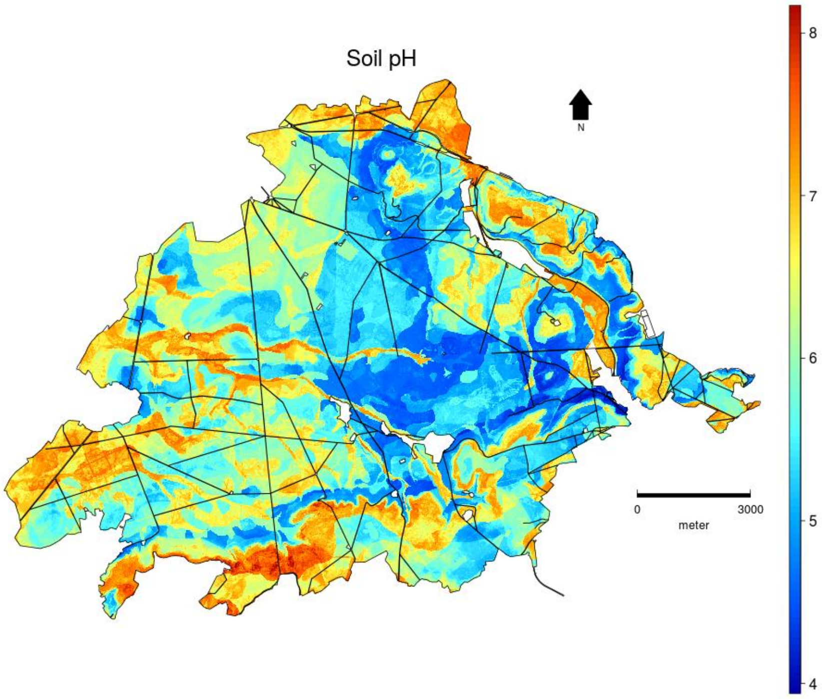
Maximum of canopy height



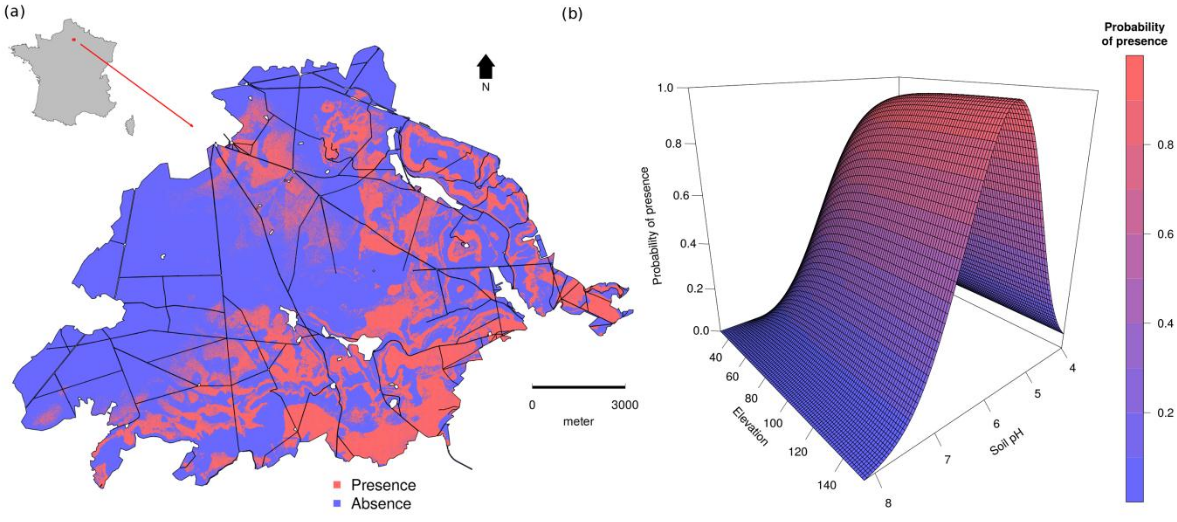
Mean of canopy height


Minimum of canopy height
N
1
2
3
4
0
3000
meter
25
20
15
10
5
0

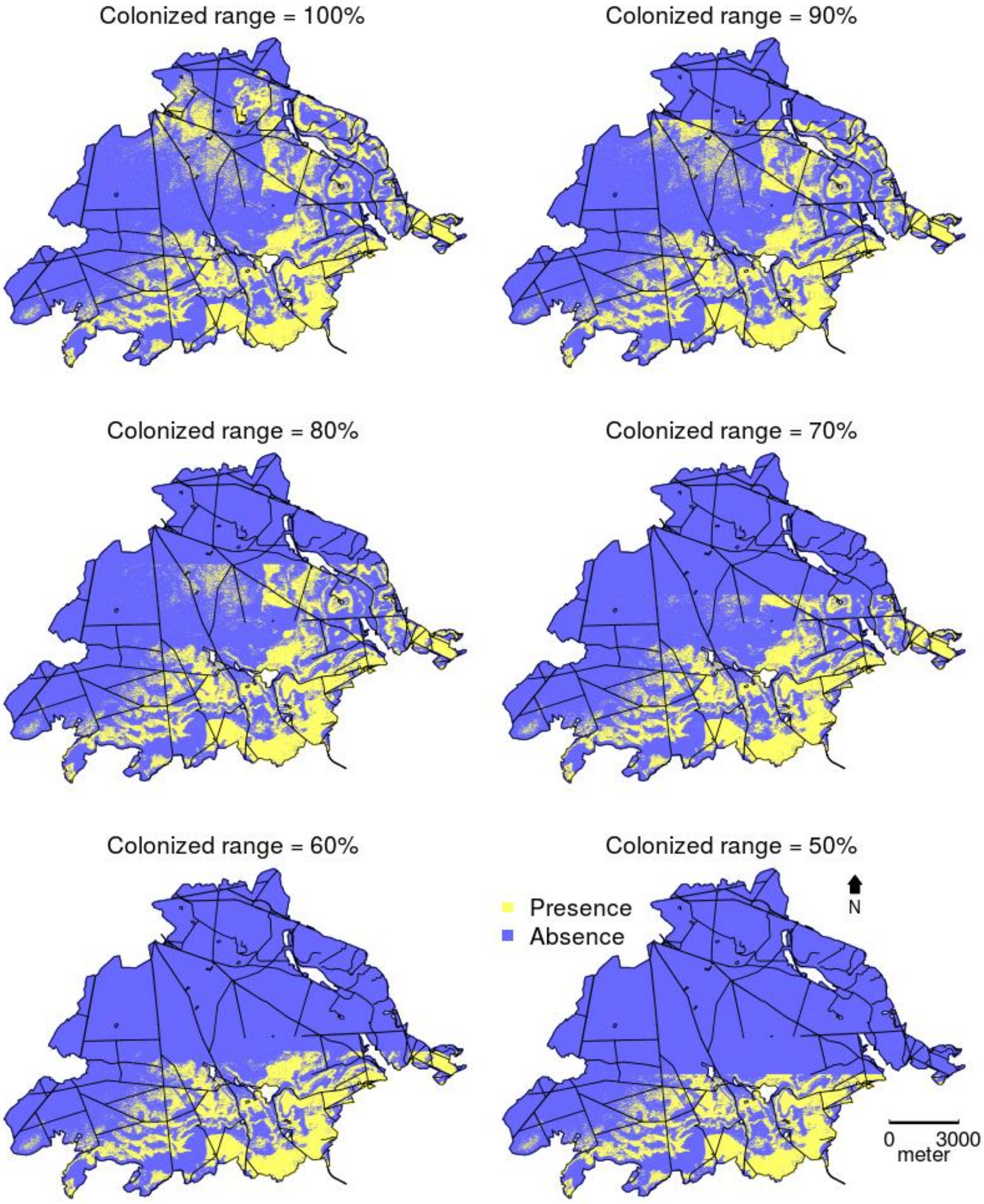

Standard deviation of canopy height
N
1
2
3
4
0
3000
meter
10
8
6
4
2
0


Soil pH
N
0
3000
meter
8
7
6
5
4

Appendix S4: Simulated (a) spatial distribution of the virtual species across the forest of Compiègne in northern France and (b) its corresponding response curve along the soil pH and elevation gradients.

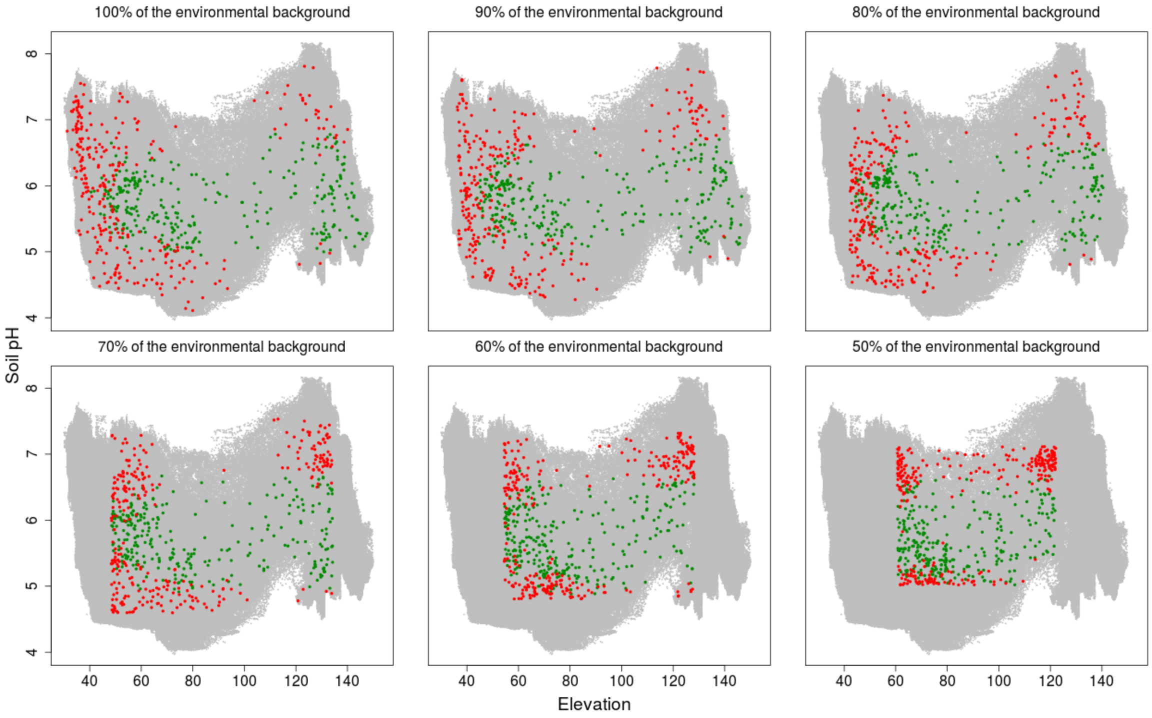


Appendix S5: Different scenarios of varying levels of alien species' invasion within its invaded range used to evaluate the performance of the computational framework



Appendix S6: Different scenarios of varying levels of environmental representativeness (cf. the proportion of the studied environmental space that is sampled by the set of presence-absence data of a given alien species within its invaded range) used to evaluate the performance of the computational framework

NB: Gray dots represent the environmental background data while red and green dots represent absences and presences, respectively.



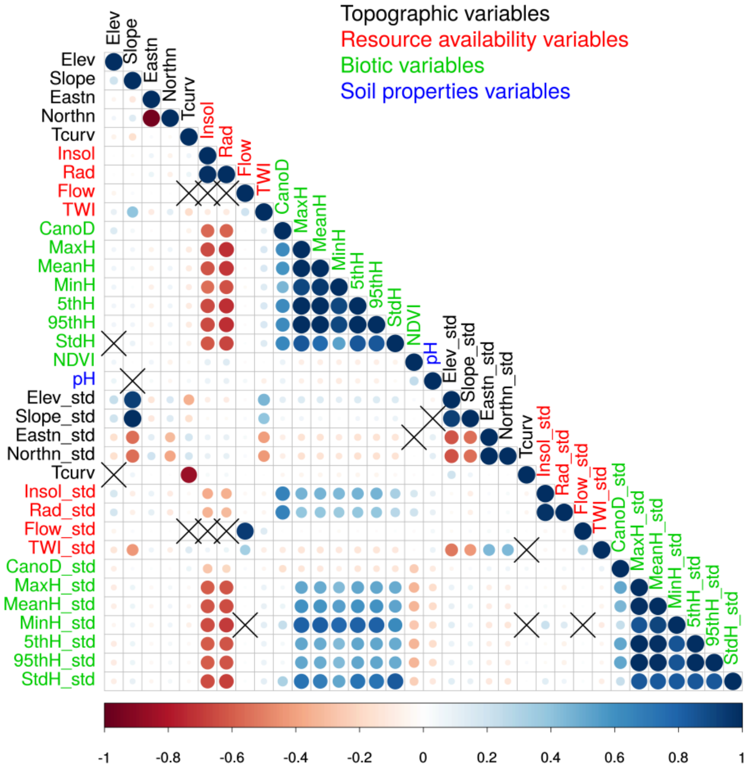
Appendix S7: Multiplicative formulation of the f λ index

$$\lambda_i = \left(1 - \frac{E_i - min(E_{ij})}{max(E_{ij}) - min(\, E_{ij})}\right) \times \frac{G_i - \min(G_{ij})}{\max(G_{ij}) - \min(\, G_{ij})}$$

where E_i is the Mahalanobis distance between the environmental conditions in the unoccupied site of interest *i* and the centroid of occupied sites within the environmental space; E_{ij} are all the Mahalanobis distances between the environmental conditions in each of the *ij* unoccupied sites and the centroids of occupied sites within the environmental space; G_i is the geographical distance between the unoccupied site of interest *i* and the nearest occupied site; and G_{ij} are all the geographical distances between each of the *ij* unoccupied sites and their respective nearest occupied sites**.**

Appendix S8: Pearson's correlation matrix showing the correlation structure among the 34 continuous predictor variables used in this study

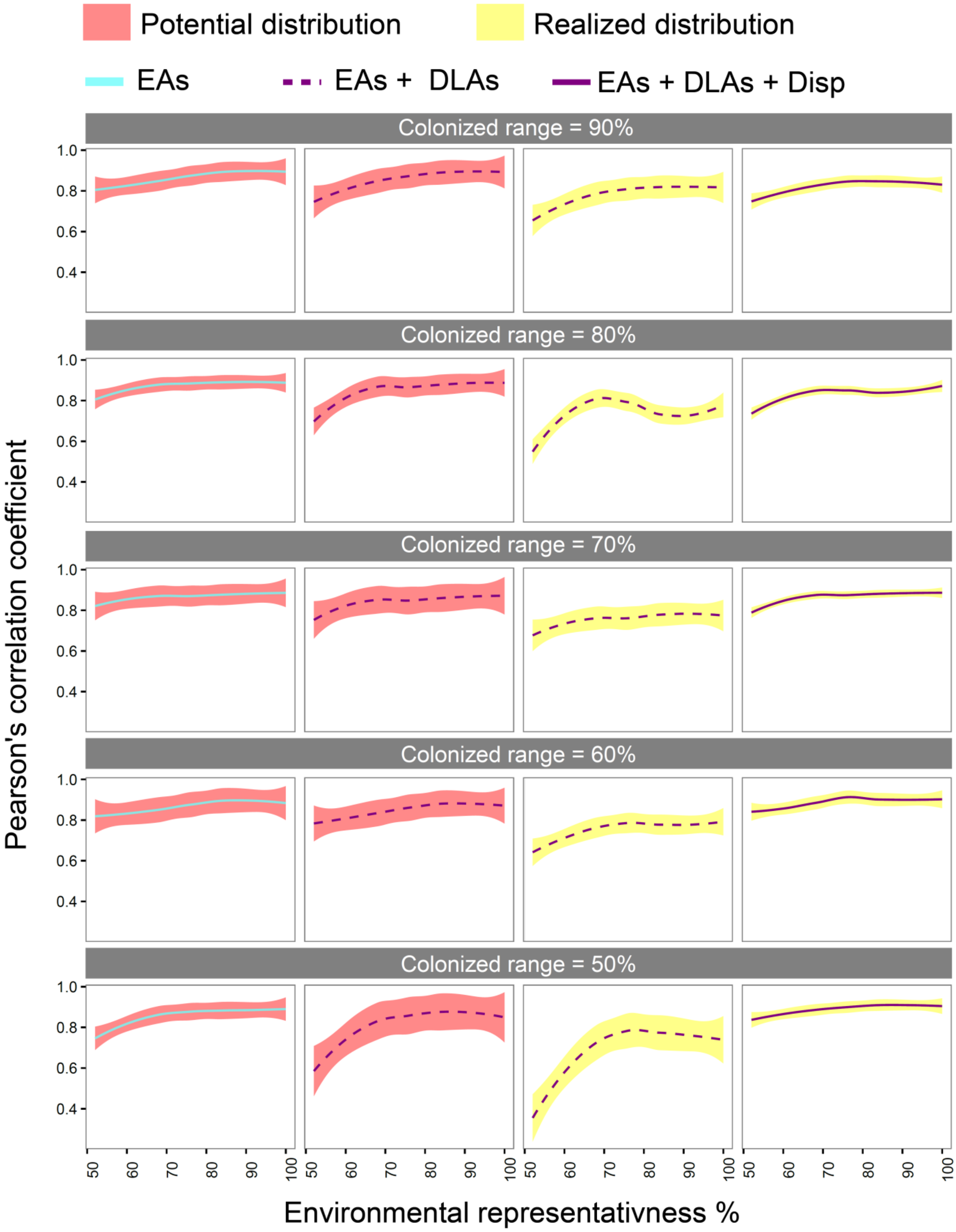
NB: Positive correlation values are displayed in blue and negative ones in red. The colour intensity and the size of the circle are proportional to the correlation value. Non-significant ($p \geq 0.05$) correlation values are displayed with a cross.



Elev = Elevation (Mean)
Slope = Slope (Mean)
Eastn = Eastness (Mean)
Northn = Northness (Mean)
Tcurv = Tangential curvature (Mean)
Insol = Insolution time (Mean)
Rad = Annuel global radiation (Mean)
Flow = Flow accumulation (Mean)
TWI = Topographic wetness index (Mean)
CanoD = Canopy density (Mean)
MaxH = Maximum height canopy (Mean)
MeanH = Mean height canopy (Mean)
MinH = Minimum height canopy (Mean)
5thH = 5th percentile height canopy (Mean)
95thH = 95th percentile height canopy (Mean)
StdH = Standard deviation height canopy (Mean)
NDVI = Normalized difference vegetation index (Mean)
pH = Soil pH (Mean)
Elev_std = Elevation (Standard deviation)
Slope_std = Slope (Standard deviation)
Eastn_std = Eastness (Standard deviation)
Northn_std = Northness (Standard deviation)
Tcurv = Tangential curvature (Standard deviation)
Insol_std = Insolution time (Standard deviation)
Rad_std = Annuel global radiation (Standard deviation)
Flow_std = Flow accumulation (Standard deviation)
TWI_std = Topographic wetness index (Standard deviation)
CanoD_std = Canopy density (Standard deviation)
MaxH_std = Maximum height canopy (Standard deviation)
MeanH_std = Mean height canopy (Standard deviation)
MinH_std = Minimum height canopy (Standard deviation)
5thH_std = 5th percentile height canopy (Standard deviation)
95thH_std = 95th percentile height canopy (Standard deviation)
StdH_std = Standard deviation height canopy (Standard deviation)

Appendix S9 Respective impacts of five different scenarios of invasion rate (see Appendix S4) and an increasing level of environmental representativeness of the set of presence-absence data (see Appendix S5) on the Pearson's correlation coefficient between predicted and simulated probabilities of presence

The shaded coloured bands and the plain/dotted lines represent the range of variability of Pearson's correlation coefficient values and the mean Pearson's correlation coefficient values, respectively, among the eight algorithms (see the materials and methods section in the main manuscript) used to model the potential and realized distributions of the virtual invasive species. For comparison purposes, we displayed results from models based on either: environmental absences (EAs: set of absences below the optimal λ threshold value; see Eq. 1 in the main manuscript) only; both environmental and dispersal-limited absences (EAs+DLAs); or all absences as well as dispersal-related covariates (EA+DLAs+Disp).


Potential distribution
Realized distribution
EAs
EAs + DLAs
EAs + DLAs + Disp
Colonized range = 90%
Colonized range = 80%
Colonized range = 70%
Colonized range = 60%
Colonized range = 50%
1.0
0.8
0.6
0.4
50
60
70
80
90
100
Pearson's correlation coefficient
Environmental representativness %