

A common benchmark for *Aedes albopictus* phenological models reveals surveillance data, not model complexity, as the limiting factor

Daniele Da Re^{1,2}, Marharyta Blaha³, Dominic Brass⁴, Cyril Caminade⁵, Miguel Garrido Zornoza⁵, Pachka Hammami⁶, Giovanni Marini¹, Mina Petric⁷, Andrea Radici⁵, Roberto Rosà³, William Wint⁸, and Kamil Erguler^{9,*}

¹Research and Innovation Centre, Fondazione Edmund Mach. San Michele all'Adige, Trento, Italy

²Georges Lemaitre Centre for Earth and Climate Research, Earth Life Institute, UCLouvain, Louvain-la-Neuve, Belgium

³Center Agriculture Food Environment, University of Trento. San Michele all'Adige, Trento, Italy

⁴UK Centre for Ecology Hydrology, Benson Lane, Wallingford, Oxfordshire, UK

⁵International Centre for Theoretical Physics, Italy

⁶Institut de Recherche pour l'Environnement, Montpellier, France

⁷Avia-GIS NV, 2980, Zoersel, Belgium

⁸Department of Biology, University of Oxford, Oxford, UK

⁹Impact and Policy Department, Climate and Atmosphere Research Centre, The Cyprus Institute, 2121 Nicosia, Cyprus

*k.erguler@cyi.ac.cy

ABSTRACT

Forecasting the seasonal dynamics of *Aedes albopictus*, a highly invasive arbovirus vector now established across much of Europe, is a growing public health priority. Numerous mechanistic and correlative models have been developed over the last two decades, but largely in isolation, each calibrated and validated against its own data, which has prevented any assessment of where predictive capacity is genuinely limited. Rather than ranking models, which were built for different purposes and cannot be reduced to a single skill score, we establish common ground: a shared evaluation framework in which six mechanistic models, one correlative machine learning model, and two naive baseline predictors are driven by identical ERA5-Land forcing and assessed against the harmonised VectAbundance ovitrap dataset (Albania, Italy, Switzerland) using a single protocol, with robustness quantified by bootstrap resampling. We additionally combine the models into a weighted ensemble via approximate Bayesian computation. Both naive baselines outperformed every process-based and machine learning model, and the ensemble matched but did not exceed the best individual models. Predictive skill varied markedly in space, with all model classes, including the uncalibrated baselines, performing best in intensively monitored regions and worst at sparsely sampled sites. This pattern suggests that geographic variation in skill reflects the density and consistency of surveillance data more than differences in model complexity. We present this benchmark not as a league table but as shared infrastructure, common evaluation data, covariates, and protocol, on which comparable, policy-relevant forecasting tools can be built, and we argue that sustained, harmonised surveillance is the principal precondition for operational *Ae. albopictus* prediction.

1 Introduction

Vector-borne diseases are expanding at an unprecedented pace, driven by the globalisation of trade and travel, rapid urbanisation, and shifting climatic conditions that are changing the geographical range of their vectors^{1,2}. Among these, *Aedes albopictus* (the Asian tiger mosquito) is a highly invasive species now established across six continents³ and a competent vector for over twenty arboviruses^{4,5}, transforming the epidemiological landscape of regions previously considered at negligible risk⁶. Europe illustrates this change: from its first recorded introduction in Albania in 1979, the species has colonised large parts of the continent, reaching as far north as Belgium, the Netherlands, and the United Kingdom, and has been implicated in a growing number of autochthonous arboviral outbreaks across Southern Europe^{2,7}.

The surge in arboviral transmission observed in Europe over the past two decades has prompted public health authorities to establish extensive surveillance and monitoring programmes^{8,9}, aimed at characterising the distribution, abundance, and seasonal dynamics of *Ae. albopictus* populations to support timely vector control. These efforts draw on a range of data sources, including egg traps (ovitrap), larval and adult mosquito surveys, and citizen reporting systems^{10–12}. Such data also underpin the modelling approaches that, alongside surveillance, play a central role in forecasting *Ae. albopictus* phenology across varying environmental and geographic contexts^{13–16}.

As ecological modellers, the authors of this study have over the past decades developed and refined models for forecasting *Ae. albopictus* abundance and phenology across large geographic scales, and several of the models evaluated here were developed by members of the author team. These models, however, have largely been built, calibrated, and validated in isolation, each against its own data. As a result, they have never been assessed on common ground, and it remains unclear how their predictive performance varies across the wide range of climatic and ecological settings the species now occupies in Europe, from Mediterranean coastal areas to continental and alpine regions. Establishing where current approaches succeed and where predictive capacity breaks down is a precondition for improving their reliability and for their application in support of public health decisions.

Structured model intercomparison has a long history in the climate and epidemiological forecasting communities, from which we draw directly. In climate projection, multi-model ensembles have been used to sample structural, parameter, and initial-condition uncertainty, while also exposing the difficulty of interpreting an “ensemble of opportunity” assembled from available rather than purpose-designed models¹⁷. The same logic has since been formalised through open, collaborative forecasting challenges for COVID-19¹⁸ and, closest to our own system, dengue^{19,20} and West Nile virus, where successive challenges systematically compared mechanistic, statistical, and ensemble approaches against naive historical baselines^{21,22}. We stress that our aim is not to produce a league table of models. The models evaluated here were built for different purposes and at different spatial and temporal scales, and no single skill metric can capture the distinct question each was designed to answer. Rather, by placing them on common ground, identical forcing data, a shared validation dataset, and a single evaluation protocol, we treat intercomparison as a diagnostic tool: a way to distinguish where predictive capacity is limited by model structure from where it is limited by the data available to constrain and evaluate it. The benchmark is thus intended less as a verdict on individual models than as shared infrastructure, a common sandbox of evaluation data and covariates, on which a community that has largely built models in isolation can begin to develop comparable, policy-relevant tools.

Specifically, this study aims to: (1) benchmark six mechanistic models, one correlative machine learning model, and two naive baseline predictors on common ground, i.e. identical ERA5-Land forcing, a single shared validation dataset (the recently published VectAbundance compilation of harmonised ovitrap egg-count observations across Albania, France, Italy and Switzerland, 2010–2022¹²), and one evaluation protocol; (2) construct and evaluate a weighted ensemble to test whether combining model outputs improves on single-model predictions; (3) characterise the geographic and environmental contexts in which these approaches succeed or fail, and assess the extent to which that variation reflects underlying surveillance data availability; and (4) establish a reusable benchmark to guide the development of more robust forecasting tools for *Ae. albopictus* surveillance and integrated vector control across Europe.

2 Methods

2.1 Entomological data

The VectAbundance dataset compiles and harmonises egg count observations of *Ae. albopictus*, collected through ovitraps as part of monitoring and surveillance efforts across Albania, France, Italy, and Switzerland from 2010 to 2022¹². In VectAbundance, the raw entomological observations were first processed into continuous time series to provide extensive geographic and temporal coverage of population dynamics, and subsequently aggregated to median values on a 9×9 km regular grid, thereby reducing the influence of minor temporal variability in monitoring.

To support model calibration and evaluation, we constructed a training dataset comprising grid cells with continuous observations over three consecutive years (2019–2021) and an independent test dataset consisting of grid cells with observations available for 2022, both spanning Albania, Italy and Switzerland. In total, 93 and 72 grid cells were included in the training

and test datasets, respectively (Figure 1a). The two datasets are temporally independent by construction, as no cell contributes observations from the same year to both. Spatially, they partially overlap: test cells coinciding with training locations (shown in purple in Figure 1a) allow us to assess temporal transferability, or interpolation, at previously observed sites, whereas test cells with no spatial counterpart in the training set allow us to assess spatial transferability, or extrapolation, to locations unseen during calibration.

2.2 Models

We utilised seven predictive models spanning mechanistic and data-driven approaches: albopictus (sPop)¹⁴, ArboCartoR¹⁶, dynamAedes¹⁵, VECTRI²³, and the models developed by Metelmann et al.¹³, Brass et al. (AedesDDE)²⁴, and Da Re et al. (Stacked Machine Learning)²⁵ (Tab. 1). Six of these are mechanistic formulations that explicitly resolve the life cycle of *Ae. albopictus*, tracking stage-specific dynamics (eggs, larvae, pupae, and adults) under environmental forcing. Despite this shared structure, the models differ substantially in their representation of biological processes, fitted thermal-dependent life-history traits functions, and environmental interactions. The albopictus(sPop) and Metelmann et al.¹³ models describe environmentally driven population dynamics based on temperature, photoperiod, precipitation, and diapause processes, whereas dynamAedes introduces stochasticity, dispersal, and multi-scale spatial structure. ArboCartoR similarly represents spatial heterogeneity through a parcel-based framework with explicit mosquito and host distributions, while VECTRI adopts a climate-driven formulation in which vector densities are coupled to temperature- and rainfall-driven processes, including a physically based representation of surface hydrology. In particular, we used the *Ae. albopictus*-parameterised version of VECTRI (v1.11.3)²⁶, in which egg, larval, and adult female densities are explicitly resolved and calibrated against long-term ovitrap observations from Emilia-Romagna, Italy. In addition to a stage-based formulation, Brass et al.²⁴ employ a size-structured model that captures phenotypic plasticity and delayed responses to environmental conditions, explicitly representing the temporal decoupling between environmental exposure and trait-mediated dynamics. Complementing these process-based approaches, the stacked machine learning model of Da Re et al.²⁵ adopts a correlative machine learning framework, combining multiple base learners to infer non-linear relationships between environmental predictors and observed egg abundance without imposing mechanistic constraints.

2.3 Ensemble model inference

The ensemble model combined predictions from the seven individual models using a weighted average,

$$\hat{y}_{\text{ens}} = \frac{\sum_{i=1}^7 w_i}{\sum_j w_j} \hat{y}_i, \quad (1)$$

where $\hat{y}_i$ is the prediction of model i and w_i its corresponding weight. Candidate weights were bounded between 0 and 1 and normalised by their sum, such that only their relative magnitudes determined each model's contribution. Combining structurally different models can improve predictive robustness by compensating for model-specific biases¹⁷.

We inferred the ensemble weights using approximate Bayesian computation (ABC). For each candidate weight configuration, the seven individual models and resulting ensemble were evaluated independently at each location using a baseline-normalised squared error,

$$D_{lm} = \frac{\sum_t (y_{lt} - \hat{y}_{lmt})^2}{\sum_t y_{lt}^2}, \quad (2)$$

where y_{lt} and $\hat{y}_{lmt}$ are the observed and predicted values, respectively. At each location, predictions were ranked according to D_{lm} , with rank 1 indicating the smallest error. The inference criterion was the mean ensemble rank across locations,

$$d(\mathbf{w}) = \frac{1}{L} \sum_{l=1}^L r_{l,\text{ens}}(\mathbf{w}), \quad (3)$$

where L is the number of locations and $r_{l,\text{ens}}(\mathbf{w})$ is the rank of the ensemble at location l for the weight configuration $\mathbf{w}$. This rank-based criterion favoured ensembles that performed consistently well relative to their constituent models while limiting the influence of differences in absolute error magnitude among locations.

We assigned each unnormalised weight an independent uniform prior over $[0, 1]$ and accepted configurations for which $d(\mathbf{w}) \leq \epsilon$. We used $\epsilon = 2.8$, corresponding to an average rank of 2.8 or better among the ensemble and its seven constituent models across all locations.

A high-density region of the ABC posterior¹⁴ was identified using an adaptive basin-hopping Markov-chain Monte Carlo algorithm implemented in the *hoppMCMC* Python package (v. 2.0.0; <https://pypi.org/project/hoppMCMC/>). We then used the *inferABC* algorithm to sample $N = 100$ weight configurations from this region using $\epsilon = 2.8$, thereby characterising uncertainty in the relative contributions of the individual models.

2.4 Baseline predictive models

To contextualise model performance, we defined two naive baseline predictive models representing minimal assumptions about temporal structure. Baseline I corresponds to a persistence model, in which egg abundance at week t is predicted as the observed value at the previous week ($t - 1$), i.e. the value is constant. This approach reflects short-term temporal autocorrelation in ovitrap observations and provides predictive skill only at a one-week horizon, thereby offering a limited window for intervention. In addition, its application requires continuous and operational surveillance data.

Baseline II represents a historical predictor, defined as the mean egg abundance for each ISO-week, a standardized seven-day period used to divide a calendar year into 52 or 53 numbered week, calculated across all available training years. This formulation captures the average seasonal cycle independent of interannual variability and climatic trends, and can be applied to generate predictions at longer lead times, including up to a full seasonal cycle in advance. Similar to Baseline I, this approach relies on the availability of consistent historical observations, but provides a broader window for potential response if seasonal patterns are sufficiently stable.

2.5 Environmental datasets

All models were driven by the same underlying environmental data: hourly gridded estimates of near-surface air temperature and total precipitation across Europe for 2019–2022, derived from ERA5-Land reanalysis²⁷. From this single source, we generated model-specific inputs by aggregating the data to the temporal resolution each model requires (e.g., daily or weekly) and formatting it to match each model's expected input structure, without altering the underlying values. This ensured that any differences in model output reflect differences in model formulation rather than differences in the forcing data.

Across models, air temperature was the only environmental covariate used consistently in all simulations, reflecting its primary role in regulating development and survival processes. Daily total precipitation was incorporated in most mechanistic models, including *albopictus*(sPop)¹⁴, *ArboCartoR*¹⁶, Metelmann et al.¹³, *VECTRI*²³, and *AedesDDE*, as well as in the stacked machine learning framework²⁵. Photoperiod was incorporated in the *albopictus*(sPop), *dynamAedes*, and Metelmann et al. models as a mechanistic cue governing diapause induction and termination, computed from latitude and day of year; it was similarly derived as a seasonal predictor in the stacked machine learning model. In these mechanistic formulations, photoperiod directly controls the start and end of the egg-laying season, so differences in its implementation are expected to affect predictions of seasonal onset and cessation. Human population density (or counts per grid cell) was incorporated in *albopictus*(sPop), *ArboCartoR*, Metelmann et al., and *VECTRI* to account for host availability or anthropogenic effects.

Additional covariates were model-specific. The *AedesDDE* model included evaporation to represent water balance effects on larval habitat, while the stacked machine learning model further incorporated land cover-based suitability indices to account for spatial heterogeneity in habitat conditions. These differences reflect the varying assumptions across models regarding the relative importance of climatic, environmental, and anthropogenic drivers, while maintaining a common core of temperature-driven dynamics.

2.6 Evaluating goodness of fit

To enable consistent comparison across models and observations, both simulated and observed time series were normalised to represent relative temporal distributions. For each observed year, weekly values were first aggregated (using the median for observations and simulations), and then scaled by the total sum over the time series, such that values represent the proportion of annual egg abundance occurring at each time step. This transformation preserves the seasonal structure of the dynamics while removing differences in absolute magnitude, allowing model performance to be evaluated based on the timing and relative intensity of activity. Cases with zero total abundance were assigned zero values.

We evaluated predictive performance using the relative squared error (RSE), defined as

$$\text{RSE} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

which corresponds to the residual sum of squares normalised by the total sum of squares^{28,29}. Here, $\hat{y}_i$ denotes the predicted value for observation y_i , and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is the sample mean of the observations. Values of RSE closer to 0 indicate better predictive performance, while $\text{RSE} = 1$ corresponds to predicting the mean. We selected an RSE threshold of 0.5 to distinguish between relatively good (lower values) and poor (higher values) fits at the grid-cell level. To compare model performance, we quantified the number of good fits per model across 100 bootstrap resamples of both the training and test datasets.

Table 1. Key structural and biological characteristics of the seven predictive models evaluated in this study. Models include six mechanistic formulations and one correlative machine learning approach, differing in their representation of environmental drivers, spatial structure, demographic detail, and underlying mathematical formalism.

Model	Mechanistic	Stochastic	Dynamics formalism	Temp.	Precip.	Photo-period	Dispersal	Habitat heterog.	Age classes	Ageing w/in class
albopictus (sPop) ^a	Yes	n.s.	n.s.	Yes	Yes	Yes	n.s.	n.s.	n.s.	n.s.
ArboCartoR ^b	Yes	n.s.	n.s.	Yes	Yes	n.s.	n.s.	Yes	n.s.	n.s.
dynamAedes ^c	Yes	Yes	n.s.	Yes	No	n.s.	Yes	Yes	n.s.	n.s.
VECTRI ^d	Yes	n.s.	n.s.	Yes	Yes	n.s.	n.s.	Yes ^e	n.s.	n.s.
Metelmann et al. ^f	Yes	n.s.	n.s.	Yes	Yes	Yes (impl.)	n.s.	n.s.	n.s.	n.s.
AedesDDE ^g	Yes	n.s.	DDE	Yes	Yes	Yes (impl.)	n.s.	n.s.	— ^h	n.s.
Stacked ML ^j	No	n/a	n/a	Yes	Yes	Yes (impl.)	n/a	Yes ^j	n/a	n/a

^a Erguler et al. (2016). Environmentally-driven, includes diapause processes.

^b Tran et al. (2020). Parcel-based framework with explicit mosquito and host distributions.

^c Da Re et al. (2022). Introduces stochasticity, dispersal, and multi-scale spatial structure; precipitation not used as a driver in this implementation.

^d Tompkins & Ermert (2013); *Ae. albopictus*-parameterised version v1.11.3 used here.

^e Via physically-based representation of surface hydrology.

^f Metelmann et al. (2019). Includes diapause processes.

^g Brass et al. (2024). Size-structured rather than discrete age-structured formulation, with delay differential equations (DDE) capturing phenotypic plasticity and delayed environmental responses; also incorporates evaporation as a water-balance covariate.

^h Size-structured rather than age-class structured; not directly comparable to discrete age classes used in other models.

ⁱ Da Re et al. (2025). Correlative stacked machine learning framework combining multiple base learners; also incorporates land-cover-based habitat suitability indices.

^j Via land-cover-based suitability indices rather than explicit spatial process representation.

n.s. = not specified in the main text; to be verified against original model publications. **n/a** = not applicable (correlative, non-mechanistic framework).

3 Results

The observational dataset is heterogeneous in spatial coverage, sampling consistency, and monitoring duration, reflecting the diversity of local surveillance schemes contributing to VectAbundance. Although the VectAbundance observations are temporally harmonised and spatially aggregated, this heterogeneity remains evident in the training and test datasets (Fig. 1a). Emilia-Romagna (Northern Italy) is the most intensively and consistently monitored region, with long-term observations from 2010–2022 and a large regional ovitrap network, whereas other regions provide shorter or more spatially restricted time series. Further characterisation of the training dataset reveals distinct regional differences in seasonal egg-laying activity, separating Emilia-Romagna from the northern provinces, while the Lazio region (Central Italy) area shows a more mixed seasonal signature (Fig. 1b). This imbalance is likely to affect model evaluation by shaping both the amount of information available for training and the geographical representativeness of the training and test datasets.

To evaluate model fits, we used an RSE threshold of 0.5, classifying fits with $\text{RSE} < 0.5$ as relatively good and all others as relatively poor. Model performance, quantified as the number of good fits across 100 bootstrap resamples, is shown for the training dataset (2019–2021) in Figure 2a and for the test dataset (2022) in Figure 2b. The two naive baselines achieved the highest number of good matches in both datasets, with Baseline I ranking first, followed by Baseline II; both outperformed all process-based and machine-learning models. By using observations from only one week earlier, Baseline I benefits from a short-term predictive advantage, although this limits its practical forecasting horizon. Baseline II, in contrast, provides a seasonal average that can be constructed from previous years and is therefore available in advance, but it applies the same average profile across years rather than capturing year-specific fluctuations. The fact that both baselines outperformed the other models under the RSE metric suggests that none of the evaluated models reproduced local fluctuations in seasonal egg-laying activity more closely than either the prior-week observation or the fixed seasonal average.

Excluding the two naive baselines, model and ensemble performance across locations revealed considerable heterogeneity in the ability to reproduce observed weekly egg-laying activity. The stacked machine-learning framework, AedesDDE, VECTRI, and the Metelmann et al. model generally performed better than albopictus(sPop), ArboCartoR, and dynamAedes in both the training and test datasets. This pattern was also reflected in the model-weight distributions used to construct the ensemble, with the better-performing models receiving higher weights more consistently (Fig. 3). We did not observe a major difference in model performance between the training and test datasets, apart from a somewhat clearer separation in the test dataset between the four better-performing models and the three poorer-performing models. The ensemble performed comparably to the average of the best individual models, but did not exceed their predictive ability.

Evaluating the ensemble prediction as a summary of overall predictive modelling capability revealed marked geographic variation in accuracy (Fig. 4). The ensemble performed best in Emilia-Romagna, where both the duration and frequency of data collection were higher than in other regions. Although this region shows a seasonal activity profile that differs from much of the remaining dataset, the sustained performance of the ensemble in the northern provinces supports its predictive capacity beyond this intensively sampled area. Sites in northern Italy and southern Switzerland generally showed good agreement with observations, reflected in low RSE values, whereas a smaller number of locations, particularly in coastal and southern areas, exhibited higher RSE values, indicating poorer fits. The best fits were concentrated in the Po Plain and Emilia-Romagna, where the monitoring network is particularly data-rich. This is expected, as the ensemble model, which combines predictions weighted by historical performance, naturally benefits from the larger volume of training observations available in these areas, allowing it to better characterise local seasonal dynamics. In contrast, data-limited locations, especially near the periphery of the study area, tended to show larger discrepancies between observations and predictions, highlighting the sensitivity of predictive ability to data availability. In Figures 4b–4f, we show representative examples spanning different levels of ensemble performance at the selected locations indicated in Figure 4a. These examples support the use of RSE as a measure of goodness of fit and further illustrate the importance of both the amount and seasonal coverage of observational data in achieving low RSE values and, therefore, good fits.

4 Discussion

This study benchmarked six mechanistic models, one correlative model and an ensemble of them against two naive baselines for predicting seasonal egg abundance of *Ae. albopictus* across multiple European sites, providing a systematic evaluation of their relative performance, transferability, and operational utility. Overall, both mechanistic and data-driven approaches, including the ensemble, performed best in data-rich areas, suggesting that spatial variation in predictive skill partly reflects uneven surveillance effort rather than genuine differences in model capacity alone. This is consistent with the broader challenge of transferability in ecological forecasting, where predictions under new conditions depend not only on model structure, but also on the quality and representativeness of the data used for calibration and evaluation^{30,31}. This is particularly relevant for *Ae. albopictus* in Europe, where harmonised datasets such as VectAbundance or Vectornet have improved the basis for abundance modelling, while remaining spatially and temporally uneven.

A central enabling factor in this study was the availability of standardised, long-term ovitrap data through the VectAbundance dataset¹². Access to observations in a consistent format across countries and years was essential not only for calibrating and benchmarking models, but also for identifying spatial patterns of model failure that would otherwise remain obscured in fragmented or nationally siloed datasets. Similar efforts are being advanced in North America by the VectorByte consortium through the VecDyn dashboard³². The concentration of best-performing fits in data-rich regions such as Emilia-Romagna, Italy, which hosts one of the longest-running and most spatially extensive standardised ovitrap surveillance networks in Europe³³, further illustrates that model reliability depends strongly on sustained observational effort. It also highlights that robust model evaluation is only possible where long-term baseline data exist.

Three features of our design temper the interpretation of this geographic pattern. First, calibration provenance is not uniform across models. The VECTRI implementation used here, for instance, was calibrated against long-term ovitrap observations from Emilia-Romagna, and other models draw on parameter estimates derived from particular regional studies. Strong ensemble performance in Emilia-Romagna therefore partly reflects a home-field advantage, as several constituent models were tuned, directly or indirectly, to this region's dynamics. Notably, however, the naive baselines, which are calibrated nowhere and depend only on local historical observations, also performed better in data-rich regions, indicating that observational density contributes to the pattern independently of model calibration. Second, while VectAbundance harmonises observations onto a common spatiotemporal grid and a consistent temporal resolution, this standardisation operates at the level of data representation rather than data generation: it cannot reconcile underlying differences in sampling effort, trap density, and servicing frequency between regional programmes, which remain heterogeneous across the dataset. Apparent underperformance in regions such as Lazio may therefore stem partly from lower or less consistent sampling rather than genuine predictive failure, with inconsistent sampling introducing noise that no model can recover. Third, disentangling these effects would require either standardisation of trapping methodology at the point of collection or an explicit observation model that bias-corrects for region-specific sampling characteristics, both of which lie beyond the scope of this study but represent important priorities for comparable future benchmarking exercises.

A further consequence of our evaluation design warrants emphasis given the operational aims of this work. Because both observed and simulated series were normalised to relative seasonal distributions, our metrics reward models that reproduce the timing and shape of the egg-laying season but are insensitive to absolute abundance. Two models that capture seasonal phenology equally well would therefore receive identical scores even if their predicted abundances differed by an order of magnitude. For the questions addressed here, anticipating the onset, peak, and decline of seasonal activity, this is appropriate, and it allows comparison across models with different absolute output scales and across sites with different trap densities. For several public health applications, however, absolute abundance is precisely the quantity of interest: the magnitude of a population peak, and not only its timing, governs biting pressure, the force of arboviral transmission, and the scale of control response required. Evaluating models on absolute abundance would demand abundance-calibrated observations and a different error structure, and remains an important direction for operationally oriented benchmarking.

A further limitation concerns the data underpinning the mechanistic models themselves. Thermal performance curves, the empirical temperature–trait relationships at the biological core of these models, are typically estimated from laboratory studies under controlled regimes, and their transferability across broad field climatic gradients cannot be assumed, so parameter uncertainty propagates into predictive uncertainty. A subtler issue is that, for several key processes, independent models often draw on the same small body of empirical work. Diapause induction is a clear example: the photoperiodic thresholds used across most models ultimately trace back to a limited set of laboratory observations³⁴, so apparent agreement between models may reflect a shared dependence on common source data rather than genuine corroboration, masking real uncertainty in how these processes vary across populations. Open-access trait databases such as VecTraits³⁵ and AedesTraits³⁶ address the parameterisation side of this problem, but adapting these relationships to field conditions remains constrained by the availability of standardised, long-term surveillance data, making this ultimately a challenge of observation as much as of model formulation.

The strong performance of naive baselines underscores that, for operational forecasts of seasonal onset or abundance in well-monitored areas, direct use of historical observations can be at least as informative as biological process representation alone. This raises a practically important question: if baselines or established individual models, such as VECTRI or Metelmann et al., already perform adequately for a given task, what does a more complex ensemble add? The answer depends critically on the purpose of prediction and on an honest appraisal of model epistemology. For temporal prediction within well-surveilled regions, simple baselines may be sufficient, provided that monitoring effort has been consistent through time. For spatial extrapolation to areas where *Ae. albopictus* is newly establishing, where surveillance remains sparse, or where climate projections extend beyond historical experience, the limitations of observation-driven approaches become more apparent. In these settings, the structured biological knowledge embedded in mechanistic models may provide value beyond empirical regularity alone. Beyond prediction under novel conditions, mechanistic models offer a capability that observation-driven approaches lack entirely: the ability to simulate counterfactual, what-if scenarios, such as the impact of alternative vector control strategies³⁷, that have no direct analogue in the historical record.

The correlative Stacked Machine Learning model also performed relatively well; however, it was trained and evaluated within the same spatial domain. Its ability to generalise beyond this domain, particularly to regions where *Ae. albopictus* is newly establishing or where surveillance data remain scarce, has yet to be rigorously assessed and should be evaluated carefully before operational deployment²⁵. This distinction is directly relevant to the design of early warning systems. For large-scale, coarse-resolution surveillance, simpler models may be entirely fit for purpose when the objective is to flag broad phenological milestones, such as the onset, peak, and offset of the mosquito season, while providing transparent, computationally tractable, and interpretable outputs.

This mirrors lessons emerging from comparable forecasting exercises in other infectious disease contexts. The COVID-19 pandemic provided an unprecedented natural experiment in large-scale ensemble modelling: despite an extraordinary volume of surveillance data and contributions from dozens of independent modelling teams, no single modelling approach consistently outperformed others, forecast skill remained highly variable across epidemic phases and geographic units, and even ensemble models struggled to anticipate abrupt regime shifts such as the Omicron wave^{18,38,39}. Critically, and directly relevant to the vector surveillance context, these failures were substantially attributable to the quality, consistency, and reporting heterogeneity of the underlying surveillance data, despite COVID-19 being one of the most intensively monitored disease events in history^{40,41}. The dengue forecasting challenges produced a convergent finding: ensemble approaches outperformed individual models not by being more mechanistically realistic, but by averaging over their respective blind spots. However, their forecast skill tended to be lowest during the periods of greatest public health relevance, when epidemic dynamics were changing most rapidly^{19,20}. A structured multi-model challenge for West Nile virus reached the same conclusion: in the CDC-organised 2020 forecasting exercise, simple models based on historical neuroinvasive disease counts generally outperformed more complex approaches, and a baseline relying on historical data alone exceeded both the ensemble and most individual team forecasts, with the inclusion of entomological surveillance data associated with lower rather than higher skill²¹. A shared lesson across all these exercises is that model selection should be governed by the question being asked and the quality of available data, not by a reflexive preference for complexity.

Recent work on this same regional dataset has begun to formalise the practical design constraints of surveillance systems in the context of machine learning: Blaha et al.⁴² demonstrated that broader spatial coverage can compensate for shorter temporal records, with configurations spanning only two recent years but incorporating more sampling locations with high revisiting frequency matching or exceeding the predictive performance of longer historical archives, a finding with direct operational implications for how new monitoring networks should be prioritised and resourced. Expanding standardised monitoring efforts into currently underrepresented areas, particularly at the northern and southern margins of the species' European range, would substantially improve the predictive capacity of both mechanistic and data-driven approaches.

We caution against reading these results as a ranking of models. Each model encodes different assumptions and was built to answer a different question, and a model that scores poorly on relative phenological timing may be well suited to a purpose this benchmark does not test, such as scenario analysis or long-range projection. The contribution of this work is therefore not a league table but a common ground: a shared evaluation dataset, a common set of environmental covariates, and a reproducible protocol against which existing and future models can be assessed on equal terms. Much vector modelling has developed in isolation, with each group calibrating and validating against its own data; establishing shared infrastructure of this kind is a precondition for cumulative progress and, ultimately, for building the kind of predictive tools that can genuinely support public health decision-making.

An operational extension follows directly. Here, models were driven by past weather time series, which allowed predictions to be compared against observations and combined into an ensemble. For an operational decision-support system able to schedule vector management in advance, however, these models would need to be coupled with medium-range weather forecasts, as is already standard in forecast-informed reservoir operation⁴³ and agricultural irrigation planning⁴⁴. Because such forecasts are updated continuously and carry their own uncertainty, this coupling calls for a continuous-integration approach that propagates forecast uncertainty into the predictions⁴⁵. Designing this system is beyond the present scope, but we view it as a natural next step, requiring collaboration between entomologists, modellers, meteorologists, and engineers to adapt existing models for operational use.

Looking ahead, we argue that the vector modelling field is now at an inflection point. For phenological prediction, i.e. anticipating the seasonal onset and peak of *Ae. albopictus* phenological activity, the current generation of models, supported by improved surveillance infrastructure and open trait databases, is already performing well. The more urgent and scientifically productive frontier lies in redirecting the structural power of mechanistic models toward questions that purely correlative or baseline approaches cannot address: evaluating intervention scenarios⁴⁶, modelling the population-level consequences of insecticide resistance, and providing decision-relevant outputs for public health stakeholders. As climate change continues to shift the thermal suitability envelope of *Ae. albopictus* northward and to higher elevations, the ability to anticipate phenological shifts in novel environments will become increasingly critical. Yet realising this potential requires confronting what is now the field's most critical bottleneck. The mathematical frameworks are available, climatic data are increasingly accessible at fine

spatial and temporal resolution, and computational capacity is no longer a limiting factor. What remains structurally lacking is harmonised, quality-controlled data sharing, not only for vector phenology, but across the full operational cycle: intervention outcomes, outbreak responses, and vector control efficacy assessments. Progress in this area is hampered not only by technical and logistical barriers, but by institutional fragmentation and the persistent treatment of publicly-funded surveillance data as proprietary assets. Coordinated investment in open, interoperable data infrastructures, governed by clear sharing agreements at the European and international level, is not a peripheral concern but a prerequisite for the next generation of operational vector surveillance tools. Without it, the capacity to move from descriptive surveillance to genuinely predictive and actionable public health science will remain fundamentally constrained, regardless of how sophisticated our models become.

References

1. WHO, G. A. I. Preparing for the next pandemic by tackling mosquito-borne viruses with epidemic and pandemic potential. *WHO: Geneva, Switz.* (2024).
2. Cattaneo, P. *et al.* Transmission of autochthonous aedes-borne arboviruses and related public health challenges in europe 2007–2023: a systematic review and secondary analysis. *The Lancet Reg. Heal.* (2025).
3. Kraemer, M. U. *et al.* The global distribution of the arbovirus vectors aedes aegypti and ae. albopictus. *elife* **4**, e08347 (2015).
4. Gratz, N. Critical review of the vector status of aedes albopictus. *Med. veterinary entomology* **18**, 215–227 (2004).
5. Gossner, C. M., Ducheyne, E. & Schaffner, F. Increased risk for autochthonous vector-borne infections transmitted by aedes albopictus in continental europe. *Eurosurveillance* **23**, 1800268 (2018).
6. Farooq, Z., Rocklöv, J. & Semenza, J. C. Northward expansion of aedes albopictus-associated arbovirus transmission risk in europe. *The Lancet Planet. Heal.* **9** (2025).
7. European Centre for Disease Prevention and Control & European Food Safety Authority. Mosquito maps (2025). [Internet].
8. Braks, M. *et al.* Vectornet: putting vectors on the map. *Front. Public Heal.* **10**, 809763 (2022).
9. Wint, G. W. *et al.* Vectornet: collaborative mapping of arthropod disease vectors in europe and surrounding areas since 2010. *Eurosurveillance* **28**, 2200666 (2023).
10. Miranda, M. Á. *et al.* Aimsurv: First pan-european harmonized surveillance of aedes invasive mosquito species of relevance for human vector-borne diseases. *Gigabyte* **2022**, 1–11 (2022).
11. Carney, R. M. *et al.* Integrating global citizen science platforms to enable next-generation surveillance of invasive and vector mosquitoes. *Insects* **13**, 675 (2022).
12. Da Re, D. *et al.* VectAbundance: a spatio-temporal database of aedes mosquitoes observations. *Sci. Data* **11**, 636 (2024).
13. Metelmann, S. *et al.* The UK's suitability for aedes albopictus in current and future climates. *J. R. Soc. Interface* **16**, 20180761 (2019).
14. Erguler, K. *et al.* Large-scale modelling of the environmentally-driven population dynamics of temperate aedes albopictus (skuse). *PLoS One* **11**, e0149282 (2016).
15. Da Re, D. *et al.* dynamaedes: a unified modelling framework for invasive aedes mosquitoes. *Parasit. Vectors* **15**, 414 (2022).
16. Tran, A. *et al.* ARBOCARTO (2020).
17. Tebaldi, C. & Knutti, R. The use of the multi-model ensemble in probabilistic climate projections. *Philos. Transactions Royal Soc. A* **365**, 2053–2075, DOI: [10.1098/rsta.2007.2076](https://doi.org/10.1098/rsta.2007.2076) (2007).
18. Chharia, A. *et al.* Accuracy of US CDC COVID-19 forecasting models. *Front. Public Heal.* **12**, 1359368, DOI: [10.3389/fpubh.2024.1359368](https://doi.org/10.3389/fpubh.2024.1359368) (2024).
19. Johansson, M. A., Apfeldorf, K. M., Dobson, S. *et al.* An open challenge to advance probabilistic forecasting for dengue epidemics. *Proc. Natl. Acad. Sci.* **116**, 24268–24274, DOI: [10.1073/pnas.1909865116](https://doi.org/10.1073/pnas.1909865116) (2019).
20. Araujo, E. C. *et al.* Leveraging probabilistic forecasts for dengue preparedness and control: The 2024 dengue forecasting sprint in brazil. *Proc. Natl. Acad. Sci.* **123**, e2508989123 (2026).
21. Holcomb, K. M. *et al.* Evaluation of an open forecasting challenge to assess skill of west nile virus neuroinvasive disease prediction. *Parasites & Vectors (Web)* **16**, 1–13 (2023).

22. Harp, R. D., Holcomb, K. M., Retkute, R. *et al.* Evaluation of the 2022 West Nile virus forecasting challenge, USA. *Parasites & Vectors* **18**, 152, DOI: [10.1186/s13071-025-06767-2](https://doi.org/10.1186/s13071-025-06767-2) (2025).
23. Tompkins, A. M. & Ermert, V. A regional-scale, high resolution dynamical malaria model that accounts for population density, climate and surface hydrology. *Malar. J.* **12**, 65 (2013).
24. Brass, D. P. *et al.* Role of vector phenotypic plasticity in disease transmission as illustrated by the spread of dengue virus by aedes albopictus. *Nat. Commun.* **15**, 7823 (2024).
25. Da Re, D. *et al.* Modelling the seasonal dynamics of aedes albopictus populations using a spatio-temporal stacked machine learning model. *Sci. Rep.* **15**, 3750 (2025).
26. Zornoza, M. G., Caminade, C. & Tompkins, A. M. The effect of climate change and temperature extremes on aedes albopictus populations: a regional case study for Italy. *J. The Royal Soc. Interface* **21**, 20240319 (2024).
27. Muñoz-Sabater, J. *et al.* ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **13**, 4349–4383 (2021).
28. Hyndman, R. J. & Koehler, A. B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **22**, 679–688 (2006).
29. Willmott, C. J. On the validation of models. *Phys. Geogr.* **2**, 184–194 (1981).
30. Wenger, S. J. & Olden, J. D. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods Ecol. Evol.* **3**, 260–267 (2012).
31. Yates, K. L. *et al.* Outstanding challenges in the transferability of ecological models. *Trends Ecol. Evol.* **33**, 790–802 (2018).
32. Rund, S. C. *et al.* Vecdyn. (2023).
33. Carrieri, M. *et al.* Surveillance of the chikungunya vector *Aedes albopictus* (Skuse) in Emilia-Romagna (northern Italy): organizational and technical aspects of a large scale monitoring system. *J. Vector Ecol.* **36**, 108–116, DOI: [10.1111/j.1948-7134.2011.00147.x](https://doi.org/10.1111/j.1948-7134.2011.00147.x) (2011).
34. Lacour, G., Chanaud, L., L'Ambert, G. & Hance, T. Seasonal synchronization of diapause phases in aedes albopictus (diptera: Culicidae). *PloS one* **10**, e0145311 (2015).
35. Johnson, L. R. *et al.* Vectraits. (2023).
36. Da Re, D. *et al.* Aedestraits: A global dataset of temperature-dependent trait responses in aedes mosquitoes. *Sci. Data* (2025).
37. Marini, G. *et al.* Effectiveness of ultra-low volume insecticide spraying to prevent dengue in a non-endemic metropolitan area of Brazil. *PLoS computational biology* **15**, e1006831 (2019).
38. Cramer, E. Y., Ray, E. L., Lopez, V. K. *et al.* Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc. Natl. Acad. Sci.* **119**, e2113561119, DOI: [10.1073/pnas.2113561119](https://doi.org/10.1073/pnas.2113561119) (2022).
39. Lopez, V. K., Cramer, E. Y., Pagano, R., Drake, J. M. *et al.* Challenges of COVID-19 case forecasting in the US, 2020–2021. *PLOS Comput. Biol.* **20**, e1011200, DOI: [10.1371/journal.pcbi.1011200](https://doi.org/10.1371/journal.pcbi.1011200) (2024).
40. Drews, M., Kumar, P., Singh, R. K. *et al.* Model-based ensembles: Lessons learned from retrospective analysis of COVID-19 infection forecasts across 10 countries. *Sci. Total. Environ.* **806**, 150639, DOI: [10.1016/j.scitotenv.2021.150639](https://doi.org/10.1016/j.scitotenv.2021.150639) (2022).
41. Sherratt, K., Gruson, H., Grah, R. *et al.* Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. *eLife* **12**, e81916, DOI: [10.7554/eLife.81916](https://doi.org/10.7554/eLife.81916) (2023).
42. Blaha, M. *et al.* Development and implementation of a passive surveillance system for aedes albopictus in Emilia-Romagna, Italy. *Ecol. Informatics* **95**, 103718, DOI: <https://doi.org/10.1016/j.ecoinf.2026.103718> (2026).
43. Delaney, C. J. *et al.* Forecast informed reservoir operations using ensemble streamflow predictions for a multipurpose reservoir in Northern California. *Water Resour. Res.* **56**, e2019WR026604, DOI: [10.1029/2019WR026604](https://doi.org/10.1029/2019WR026604) (2020).
44. Simionesei, L., Ramos, T. B., Palma, J., Oliveira, A. R. & Neves, R. IrrigaSys: A web-based irrigation decision support system based on open source data and technology. *Comput. Electron. Agric.* **178**, 105822, DOI: [10.1016/j.compag.2020.105822](https://doi.org/10.1016/j.compag.2020.105822) (2020).
45. Agüera-Pérez, A., Palomares-Salas, J. C., González de la Rosa, J. J. & Florencias-Oliveros, O. Weather forecasts for microgrid energy management: Review, discussion and recommendations. *Appl. Energy* **228**, 265–278, DOI: [10.1016/j.apenergy.2018.06.087](https://doi.org/10.1016/j.apenergy.2018.06.087) (2018).

46. Haramboure, M. *et al.* Modelling the impact of sterile male releases on a wild mosquito population–model assessment from field trials in mauritius. *Peer Community J.* **6** (2026).

Acknowledgements

Daniele Da Re was supported by the Marie Skłodowska-Curie Actions - Postdoctoral fellowship Nr. 101106664. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission. Kamil Erguler received funding from the Wellcome Trust (226065/Z/22/Z). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions statement

Conceptualisation: Daniele Da Re, Kamil Erguler and Roberto Rosà conceived the initial idea for the study, with relevant contributions from all the co-authors. *Data analysis:* Daniele Da Re, Kamil Erguler, Marharyta Blaha, Dominic Brass, Cyril Caminade, Pachka Hammami, Miguel Garrido-Zornoza, and Andrea Radici performed the exploratory analysis to tune the single models to the common data infrastructure. Kamil Erguler, with contributions from Daniele Da Re, developed the ensemble modelling pipeline and evaluation. *Writing:* Daniele Da Re and Kamil Erguler led the writing of the first manuscript draft. All authors contributed critically to the development, revision, and final approval of the manuscript.

Additional information

This manuscript is under active revision ahead of journal submission; a revised version will be posted.

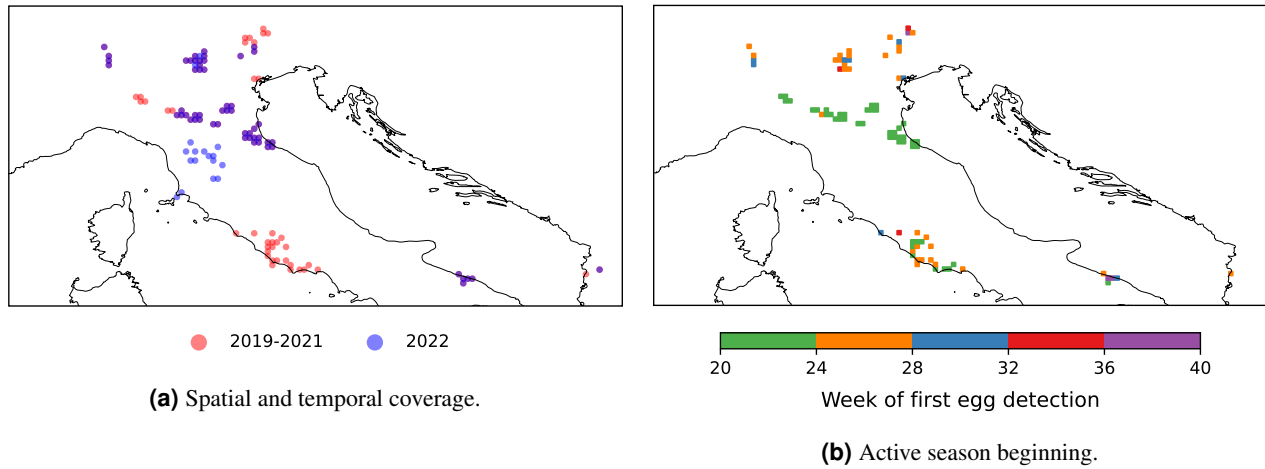
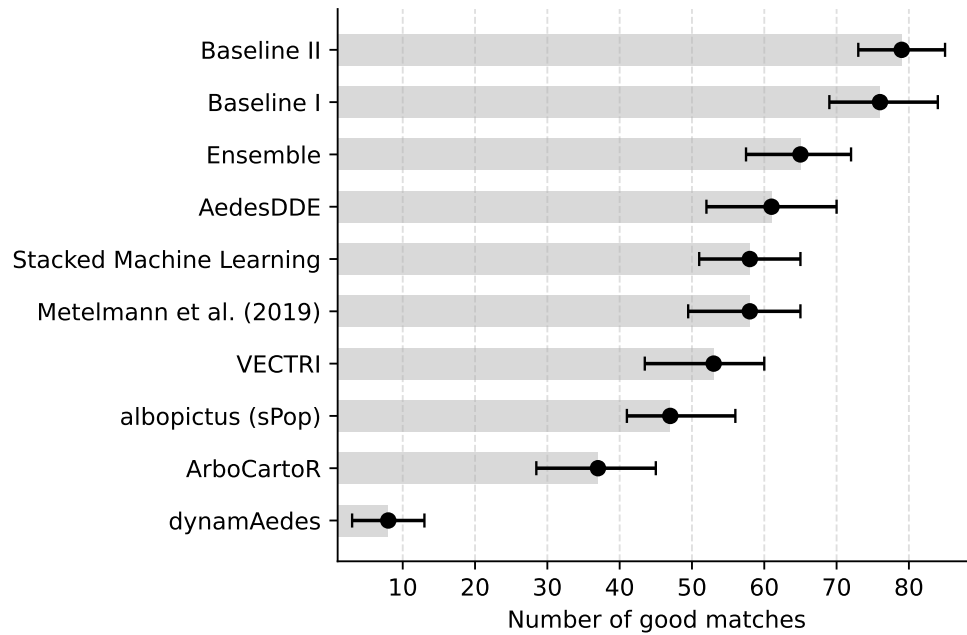
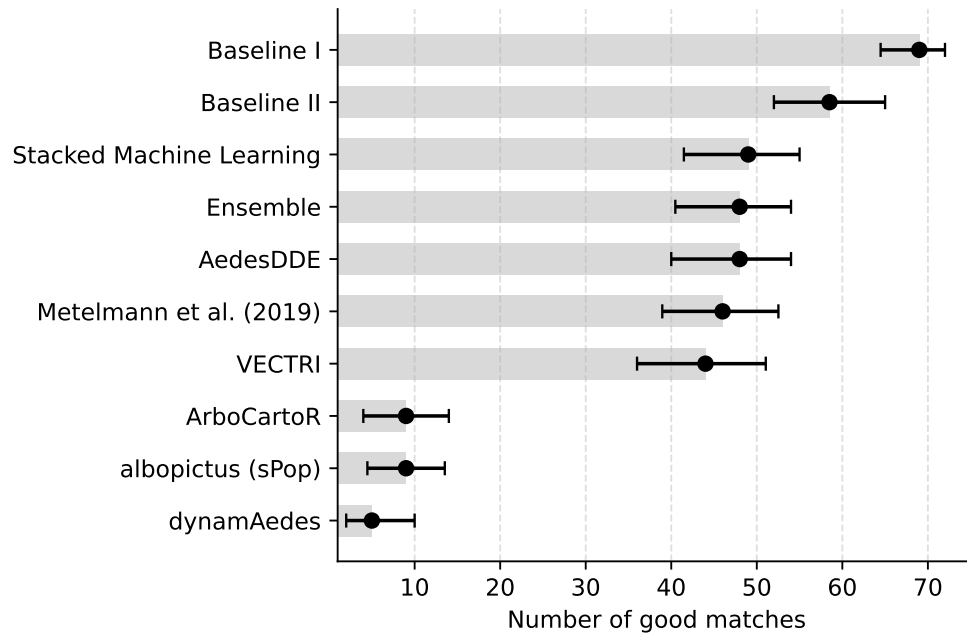


Figure 1. Spatiotemporal profiling of the VectAbundance dataset. (a) Spatial and temporal coverage showing locations with observed *Ae. albopictus* eggs between 2019–2021 (red) and in 2022 (blue). Overlapping locations are indicated with purple dots. (b) Environmental similarity of the locations observed between 2019–2021, measured by the week of first egg detection.



(a) The number of good fits in the training dataset.



(b) The number of good fits in the test dataset.

Figure 2. Evaluating the goodness of fit. Comparison of predictive performance across the training (a) and test (b) datasets are shown. In panels (a) and (b), the median and 95% range of the number of good fits (grid cells with $RSE < 0.5$) are shown for each model across 100 bootstrap resamples.

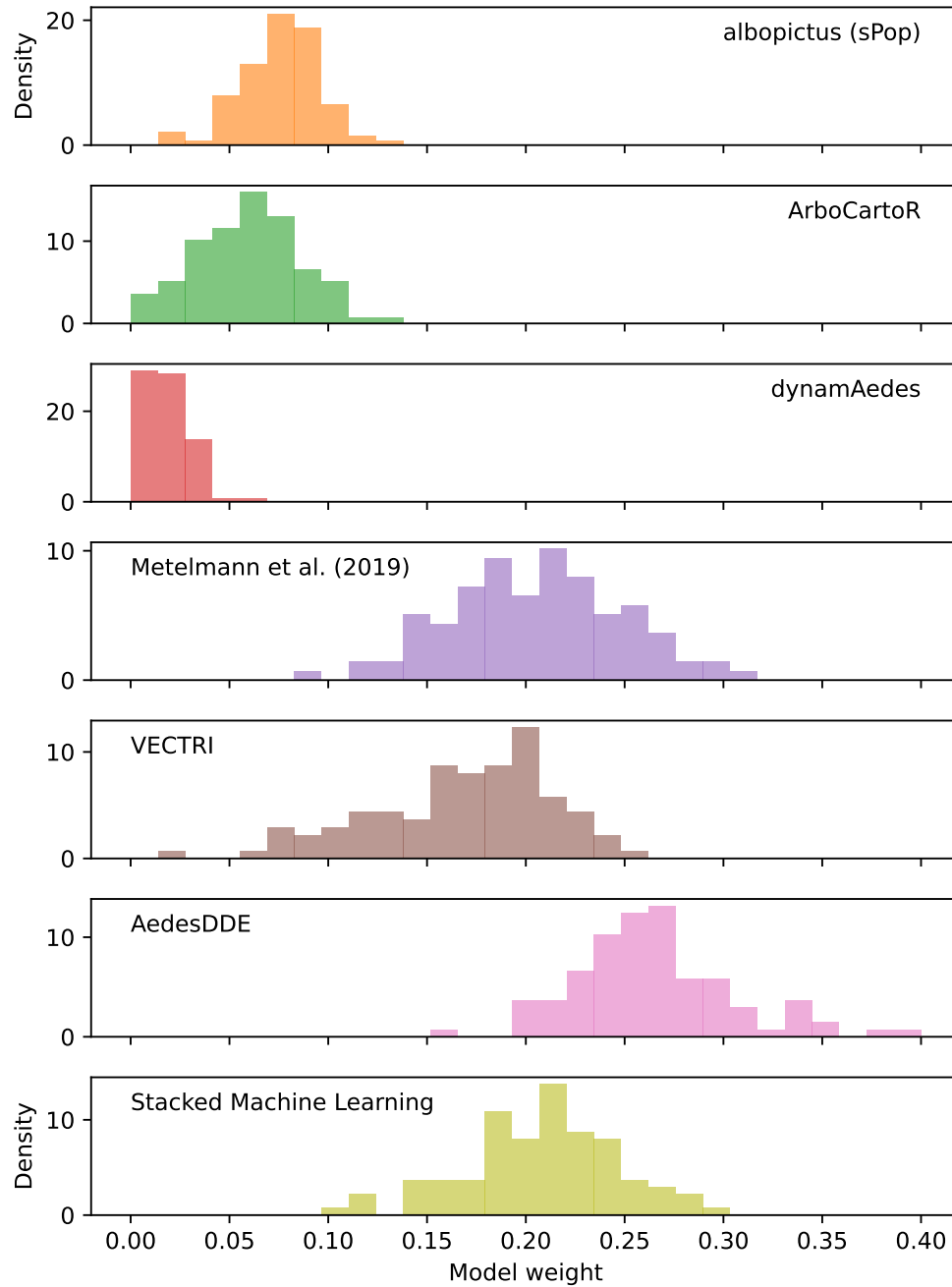
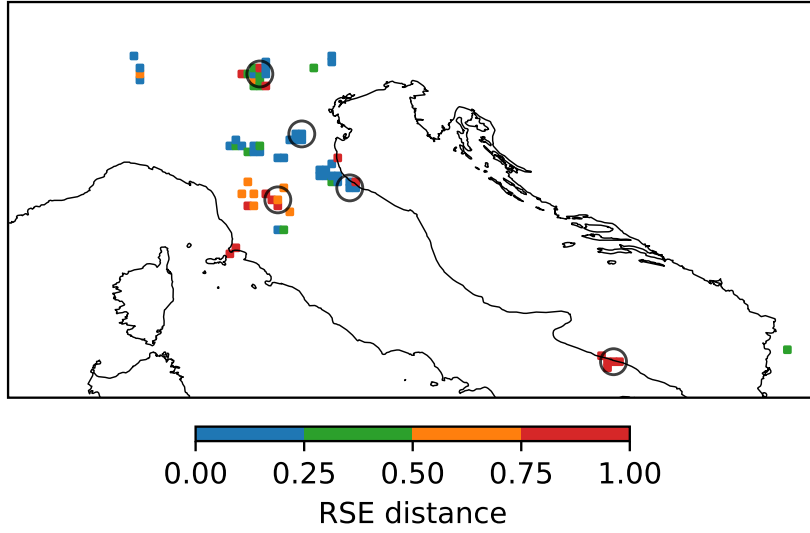
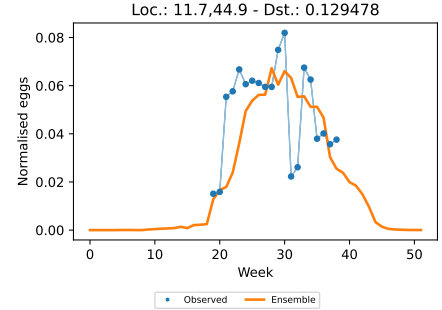


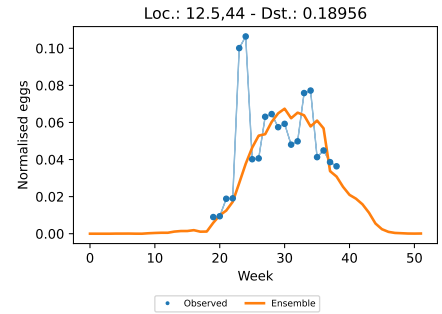
Figure 3. Distribution of model weights in the ensemble model.



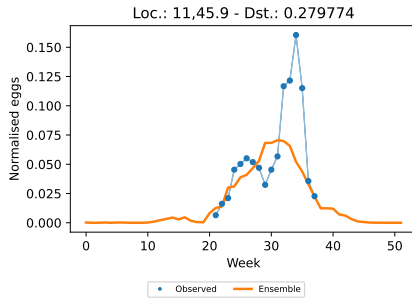
(a) Spatial goodness of fit.



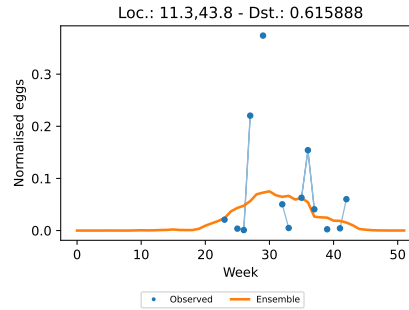
(b) Caption for panel b.



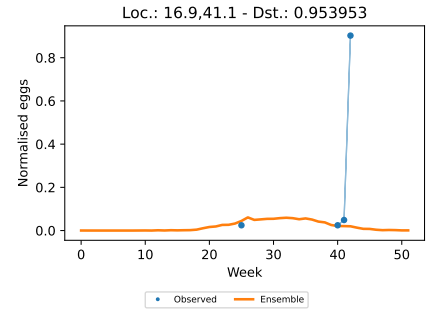
(c) Caption for panel c.



(d) Caption for panel d.



(e) Caption for panel e.



(f) Caption for panel f.

Figure 4. Spatial and temporal evaluation of ensemble model performance. Panel (a) shows the spatial distribution of predictive performance across the 2022 test dataset, with each grid cell coloured according to the mean RSE of the ensemble model. Panels (b)–(f), marked in panel (a), present weekly time series of observed (blue dots) and ensemble-simulated (orange lines) egg counts for selected locations, normalised as described in the Methods, ordered from best (b) to poorest (f) fit. Lower RSE corresponds to closer agreement between observations and simulations; the progression illustrates how fit quality declines as observational coverage becomes sparser and more irregular