How birds weather the weather: avian migration in the mid-latitudes

Kemp, M.U.

Publication date
2012

Citation for published version (APA):
Predicting migration intensity at multiple sites using operational weather radar

6.1 Introduction

The first reported bird strike to a powered aircraft was recorded in the journal of the first pilot on record: Orville Wright (Thorpe 2003). It should not be surprising then that bird strikes remain a significant threat to aviation still today (Allan 2002; Dolbeer et al. 2000). Between 1950 and 1999, at least 190 European military aircraft were lost due to birds (Richardson and West 2000). In the United States alone, the Federal Aviation Administration (FAA) reports that over 100,000 bird strikes to general and civil aircraft were reported (voluntarily) between 1990 and 2010 (Dolbeer et al. 2012).

In civil aviation, flights often cannot be canceled or even significantly delayed without incurring large financial consequences. It is often impractical, therefore, to avoid bird strikes by restricting these types of flights for any considerable length of time. Rather, these industries have focused their efforts on tactics such as land management (i.e. habitat modification), harassment or scaring techniques, and species removal. More than 70% of the bird strikes in the FAA’s report occurred below 500 ft. (∼152 m) above ground level and decreased exponentially thereafter, suggesting that the risk of a bird strike in civil and general aviation is concentrated primarily around aerodromes during take-off and landing. While these are critical phases of flight, they comprise a relatively small percentage of the total flight time. Military training flights, on the other hand, frequently occur at lower altitudes where the risk of a bird strike can be high for the entirety of the flight; however, for military training flights, adjusting flight trajectories, delaying take-off, or even cancelling flights is both practical and feasible (Shamoun-Baranes et al. 2008; van Belle et al.)
Because military aviation has the option to adjust, delay, or cancel training flights at times of intense migratory activity, countries have developed systems to monitor (near) real-time migratory activity to make these determinations and to predict the intensity of migration from environmental variables including forecasted weather conditions (e.g., van Belle et al., 2007; Rabøl, 1974; Blokpoel, 1969; Bouten et al., 2005, 2003). Predictive models of bird migration intensity have primarily been calibrated using localized measurements of migration intensity obtained from direct observation, infrared devices, military surveillance radar, or dedicated bird-detection radar; however, models calibrated for one location do not necessarily perform well at other locations far away (van Belle et al., 2007), so the predictions from these models may have a limited spatial range in which they are valid.

In this study, we aim to develop an ensemble of models to forecast migration intensity that is calibrated using measurements of migration intensity obtained from existing operational weather radar at two locations in the Netherlands. Because models calibrated for one location do not necessarily perform well far away from that location, the use of operational weather radar in the development of these models is particularly attractive. Vast networks of operational weather radars are already in place – e.g., the radar systems involved in the Operational Programme for the Exchange of weather RAdar information or OPERA network in Europe (Holleman et al., 2008), the Baltic Sea Experiment radar network (BALTRAD; Alestalo, 2002), and the Next-Generation Radar or NEXRAD network in the United States (Chilson et al., 2012) – creating the potential for standardized locally-calibrated predictive models covering enormous geographic areas. Thus, a second aim of this study is to outline a model-development procedure that can be easily applied to new locations. The models developed in this study, and the model-development procedure described, are intended to produce the most accurate predictions of migration intensity possible, which requires a different approach from the development of models to better understand the relationship between migratory dynamics and environmental conditions. Because we intend for this model-development procedure to be robust and generally applicable, we employ generalized additive models (GAMS; Hastie and Tibshirani, 1990) in development and testing. GAMS are particularly useful in this context because they are not constrained to predefined parametric relationships between predictor and response variables.
Figure 6.1: At left, a map of the Netherlands with large black circles indicating the ranges of the C-band Doppler weather radar sites in De Bilt (52.11°N 5.18°E) and Den Helder (52.95°N 4.79°E). White triangles indicate the closest grid point in the HIRLAM dataset to each radar, from which most weather data were obtained. A white square indicates the location of the NCEP grid point from which temperature data were obtained to calculate climatological normals. At right, a vertical profile of the radar measurement volume indicating the 18 altitude bins within which measurements of bird density ($B_d$) were calculated.

6.2 Materials

6.2.1 Radar measurements of migration intensity

From the spring (1 February - 31 May) and autumn (1 August - 30 November) of 2008 and 2009, we used methods described by Dokter et al. (2011) to derive altitude profiles of bird density ($B_d$; birds/km$^3$) and average speed ($B_{spd}$; m/s$^{-1}$) and direction ($B_{dir}$; ° clockwise from north) relative to the ground every five minutes from two C-band Doppler weather-radars located in the Netherlands (De Bilt 52.11°N 5.18°E and Den Helder 52.95°N 4.79°E; see Figure 6.1). Each altitude profile described $B_d$, $B_{spd}$, and $B_{dir}$ from 0.4 to 4 km above the ground in altitude bins of 200 m. Thus each profile consisted of 18 measurements, each calculated from within a circular measurement window extending from 5 to 25 km laterally from the center of the radar (Figure 6.1).

As a means of additional quality control, we used HIRLAM wind data (see section 6.2.2) to calculate airspeeds from $B_{spd}$ and $B_{dir}$ by vector subtraction. We set $B_d$ measurements to zero if the associated airspeed was not between
7 and 25 ms\(^{-1}\), as this range captures the airspeeds of the majority of avian migrants\cite{Bloch1982, Bruderer2001} and largely excludes the airspeeds of migrating insects\cite{Alerstam2011, Aralimarad2011, Chapman2008}.

We calculated height-integrated bird density (iBd; birds/km\(^2\)) from each 5-minute interval altitude profile as

\[iBd = \sum_{h=1}^{18} Bd_h \cdot \Delta h\]  \hspace{1cm} (6.1)

where \(Bd\) was integrated over the 18 altitude bins \((h)\). We then aggregated these 5-minute interval \(iBd\) measurements into one-hour averages to serve as the response variable in our models.

### 6.2.2 Variables used to predict hourly migration density

#### Baseline intensity

Migratory activity is known to exhibit seasonal and diel dynamics that should be represented in models to predict migration intensity. One option is for researchers to calculate baseline intensity explicitly as the average bird density in their dataset having occurred at a particular time of the day and year\cite{vanBelle2007}. In linear and generalized linear models, this method has the advantage of being able to capture a potentially non-linear process over time using a variable that can be treated linearly in these models. Through the use of smoothing terms, GAMs are able to capture non-linear processes directly in a model. This allows for the use of time directly in our models, permitting the GAM fitting procedure to determine the optimal (possibly non-linear) representation of the influence of time on migration intensity. Thus, we use the day of the year (DOY) and the proportion of the day (\(P_{day}\)) or night (\(P_{night}\)) directly in our models to reflect any temporal patterns in migration intensity. \(P_{day}\) and \(P_{night}\) were calculated relative to sunrise and sunset such that \(P_{day} = 0\) at sunrise and 1 at sunset and, conversely, \(P_{night} = 0\) at sunset and 1 at sunrise.

The use of GAMs also allowed for the incorporation of interaction terms: that is, terms which account for the fact that the influence of one predictor variable on the response is dependent upon the value of the other predictor variable. Thus, the two predictor variables ‘interact’ with one another regarding their influence on the response variable. We expected that DOY and \(P_{day}\) or \(P_{night}\) would likely exhibit such a dynamic relationship, because, for example, the times of sunrise and sunset – and therefore the lengths of day and night – were variable through a migration season. We therefore included these
time-related variables as interaction terms in our models. Because the units of \(DOY\) and either \(P_{\text{day}}\) or \(P_{\text{night}}\) were quite different from one another, we used non-isotropic tensor splines (Wood, 2006) to fit their functional relationship to \(iBd\) in our models. Hereafter, we refer to this smoothed interaction term between \(DOY\) and either \(P_{\text{day}}\) or \(P_{\text{night}}\) as a single variable called \(s\text{Time}\). This baseline intensity variable \(s\text{Time}\) was included in all models by default.

**Autocorrelation**

Erni et al. (2002b) suggested that migration intensity was autocorrelated on successive nights; however, Erni et al. (2002b) did not include a parameter to explicitly account for this autocorrelation in their analysis because for their analysis they desired that “predictions for migration should depend only on weather conditions and the date and not on previous observations, which may not exist”. Our aim in this study was to develop forecast models of migration intensity for flight safety, so incorporating previous measurements of bird density to improve the accuracy of our predictions was desirable. We therefore measured the autocorrelation between successive \(iBd\) measurements in our dataset, and included two variables meant to capture temporal autocorrelation in our models: 1) the mean \(iBd\) value over the previous hour (hereafter \(iBd_h\)) and 2) the mean \(iBd\) value on the previous day or night (hereafter \(iBd_d\)).

**Atmospheric variables**

To aid in the development of future models, we relied exclusively on atmospheric data available in the High Resolution Limited Area Model (i.e. HIRLAM; Cats and Wolters 1996; Undén et al. 2002) or the freely-available National Centers for Environmental Prediction (NCEP)/Department of Energy (DOE) Reanalysis II dataset (Kanamitsu et al. 2002) (see Chapter 2 for more details). HIRLAM, from which the majority of variables were obtained, is a high-resolution gridded atmospheric model that reflects the combined initiatives of the meteorological offices of multiple European countries to develop and maintain a numerical short-range weather forecasting system for operational use (Cats and Wolters 1996; Undén et al. 2002). Our use of data from this model has several likely benefits: 1) for practicality, the model-development procedure we apply will be directly exportable to new locations in Europe and elsewhere that HIRLAM data are available – importantly, the organizations developing and implementing the HIRLAM model are often the same organizations operating and maintaining the weather radar systems being used to quantify bird migration; 2) because HIRLAM is a forecasting
system, models resulting from this development procedure may be applied to-ward the prediction of migration intensity to serve in flight safety and other contexts such as mitigating bird-strikes with wind turbines (Desholm et al., 2006); 3) predictions of migration intensity can potentially be made for locations without a weather radar using local weather conditions obtained from HIRLAM as input into models of migration intensity calibrated for nearby locations.

Using data from the gridded HIRLAM atmospheric model, we derived altitude profiles, to a height of 1 km, of wind condition ($ms^{-1}$), temperature ($T$; $K$), atmospheric pressure ($AP$; mb), and relative humidity ($RH$; %). These data had a spatial resolution of $0.1\degree \times 0.1\degree$ on a rotated grid, temporal resolution of one hour, and were discretized vertically at fixed pressure levels separated by not more than 12 mb. Using data from the HIRLAM grid point nearest the center of the De Bilt radar ($\sim$33 km east at 52.02\degree N 5.64\degree E; see Figure 6.1) or the Den Helder radar ($\sim$41 km south at 52.59\degree N 4.91\degree E; see Figure 6.1), we calculated averages of each variable from these vertical profiles. We then calculated the 24-hour change in $T$, $AP$, and $RH$, hereafter denoted $\Delta T$, $\Delta AP$, and $\Delta RH$, respectively. We also derived estimates of the accumulation of precipitation ($R$; mm) over each hour from the HIRLAM model.

Wind data were described by two components, $U$ and $V$ ($ms^{-1}$), indicating the speed and direction into which the wind was blowing. The $U$ vector described the wind’s east/west component (toward east being positive) and $V$ described the north/south component (toward north being positive). Calculating flow-assistance is a useful way to reduce the complex and non-linear effects of the two components of a flow (e.g. $U$ and $V$) into a single value that facilitates quantitative comparisons between different flow conditions and incorporation of wind support into linear models (see Chapter 3); however, GAMs allow for the inclusion of interaction terms in which the influence of each variable is dependent on the value of the other. Thus, we can include $U$ and $V$ wind components in our models directly as interaction terms and avoid any potential loss of information associated with calculating flow-assistance (e.g. arising from the assumption of a single preferred direction of migration). Hereafter, we refer to this smoothed interaction term composed of $U$ and $V$ wind components as $sWind$. Note that we tested several of the representations of wind profit described in Chapter 3 and found that $sWind$ always resulted in more accurate predictions.

Following Erni et al. (2002b), we calculated rain duration ($Rdur$) in thirds for each day and night, describing the proportion of hours of the day or night that $R$ was $>0$, and applied that value to all observations during the asso-
6.3. METHODS

Migration can be particularly intense following successive days/nights with unsupportive weather conditions (Richardson 1990a). Therefore, again following Erni et al. (2002b), we calculated the potential accumulation of migrants due to previous days/nights with precipitation as 

$$ R_{acc_i} = \frac{1}{3} R_{acc_{i-1}} + \frac{2}{3} rain_{i-1} $$

where $rain$ was set to zero for days/nights without rain ($R_{dur} = 0$) and one for days/nights with rain ($R_{dur} > 0$). Similarly, we calculated an accumulation effect due to successive days/nights with unsupportive wind conditions as 

$$ W_{acc_i} = \frac{1}{3} W_{acc_{i-1}} + \frac{2}{3} wind_{i-1} $$

where $wind$ was set to zero for days/nights with supportive winds and one for days/nights with unsupportive winds. To define supportive and unsupportive winds, we used $EQ_{Tailwind}$ (see Section 3.3.1) with the preferred direction of migration set to the circular mean (Batschelet 1981) of the measured tracks (i.e. $Bdir$) and calculated headings for the particular season, location, and time (i.e. diurnal or nocturnal). We considered wind conditions supportive if $EQ_{Tailwind} > -7$ during the first hour of the day or night and unsupportive if $EQ_{Tailwind} \leq -7$ during the first hour of the day or night. The threshold of -7 for acceptable winds was also taken from Erni et al. (2002b). We then calculated a combined cumulative effect of successive days/nights with rain or unsupportive winds as 

$$ RW_{acc_i} = \frac{1}{3} RW_{acc_{i-1}} + \frac{2}{3} wx_{i-1} $$

where $wx$ was set to zero for days/nights without rain and with supportive winds and one for days/nights with rain or unsupportive winds.

Richardson (1990a) noted that, in order to separate the effect of temperature from seasonal and diel effects on migration intensity, deviations from the average temperature for the time of day and year should be used rather than absolute temperatures. We therefore used the RNCEP package (see Chapter 2) to calculate climatological temperature normals using data from the National Centers for Environmental Prediction (NCEP)/Department of Energy (DOE) Reanalysis II dataset (hereafter referred to as R-2; Kanamitsu et al. 2002). We obtained temperature data ($K$) from the R-2 grid point closest to the De Bilt and Den Helder radar sites (located at 52.5°N 5°E; see Figure 6.1) for the 1000 and 925 mb pressure levels in six-hour intervals from 1980-2010. Using the mean of the temperatures in the two pressure levels at each time step, we calculated temperature normals ($T_{norm}$) per day of year and time of day using tensor product smooths (Wood 2006). We calculated deviations from these normals as $T - T_{norm}$ and refer to these deviations hereafter as $T_{dev}$.

### 6.3 Methods

Based on a quasi-Poisson distribution, we applied penalized likelihood fitting to estimate the smoothness of smoothed terms in our GAMs. Computations
were done in the R language (R Development Core Team, 2010) using the \texttt{gam()} function from the mgcv package (Wood, 2008). In all cases, we calibrated models for diurnal and nocturnal migration independently for each radar and season. Table 6.1 gives an overview of the predictor variables considered in these analyses.

Throughout these analyses, we refer to model ‘performance’, which was determined by 50-times repeated random-sampling cross-validation using 70% of data for calibration, leaving 30% for testing. A model’s performance was the mean absolute deviation ($MAD$) between the model predictions and the 30% of data set aside for testing averaged over all 50 cross-validation iterations.

6.3.1 Model development

Benchmark models

We developed four benchmark models and determined the performance of each. Benchmark models only contained variables to account for the time of day and year and autocorrelation and did not contain variables to account for atmospheric dynamics. One benchmark model contained only $sTime$; a second was composed of $sTime + iBd_h$; a third was composed of $sTime + iBd_d$; a fourth was composed of $sTime + iBd_h + iBd_d$.

Ensemble model development

From a base model containing only the variable $sTime$, we developed an ensemble of models in which one model was calibrated for each possible combination of predictor variables, considering up to five predictor variables (plus $sTime$) in any one model. Ultimately, these models were calibrated using all available data; however, we first determined the performance of each model according to our cross-validation procedure. Models with better performance (i.e. lower $MAD$) were given more weight in the resulting ensemble. The weight ($WGT$) of a prediction ($Pred$) for the $i^{th}$ model in the ensemble was defined as

$$WGT_i = \frac{MAD_{max} - MAD_i}{MAD_{max} - MAD_{min}},$$

(6.2)

where $MAD_{max}$ was the largest and $MAD_{min}$ the smallest $MAD$ value of any of the $n$ models in the ensemble. An ensemble prediction ($Pred_E$) was then calculated as the weighted mean of the individual predictions of the $n$ models in the ensemble as

$$Pred_E = \sum_{i=1}^{n} WGT_i \cdot Pred_i.$$  

(6.3)
Table 6.1: Variables used to predict hourly migration intensity ($iBd$). The first column provides the abbreviated form of the variable used in the text; the second column indicates the units of the variable; and the third column provides a brief description of the variable.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$sTime$</td>
<td>–</td>
<td>Smoothed interaction term composed of the day of the year (DOY) and the proportion of the day ($P_{day}$) or night ($P_{night}$)</td>
</tr>
<tr>
<td>$iBd_h$</td>
<td>$\ln(\frac{\text{birds}}{\text{km}^2})$</td>
<td>Mean $iBd$ the previous hour</td>
</tr>
<tr>
<td>$iBd_d$</td>
<td>$\ln(\frac{\text{birds}}{\text{km}^2})$</td>
<td>Mean $iBd$ the previous day or night</td>
</tr>
<tr>
<td>$sWind$</td>
<td>ms$^{-1}$</td>
<td>Smoothed interaction term composed of $U$ (east/west) and $V$ (north/south) wind components</td>
</tr>
<tr>
<td>$Tdev$</td>
<td>K</td>
<td>Deviation from normal of the average temperature between 0.4 to 1 km</td>
</tr>
<tr>
<td>$AP$</td>
<td>mb</td>
<td>Average atmospheric pressure between 0.4 to 1 km</td>
</tr>
<tr>
<td>$RH$</td>
<td>%</td>
<td>Average relative humidity between 0.4 to 1 km</td>
</tr>
<tr>
<td>$\Delta T$</td>
<td>K</td>
<td>24-hour change in temperature</td>
</tr>
<tr>
<td>$\Delta AP$</td>
<td>mb</td>
<td>24-hour change in $AP$</td>
</tr>
<tr>
<td>$\Delta RH$</td>
<td>%</td>
<td>24-hour change in $RH$</td>
</tr>
<tr>
<td>$R$</td>
<td>mm</td>
<td>Precipitation amount</td>
</tr>
<tr>
<td>$Rdur$</td>
<td>thirds</td>
<td>Duration of precipitation over a day or night</td>
</tr>
<tr>
<td>$Racc$</td>
<td>–</td>
<td>Accumulation of migrants due to precipitation</td>
</tr>
<tr>
<td>$Wacc$</td>
<td>–</td>
<td>Accumulation of migrants due to unsupportive winds</td>
</tr>
<tr>
<td>$RWacc$</td>
<td>–</td>
<td>Accumulation due to precipitation or unsupportive winds</td>
</tr>
</tbody>
</table>
We indicated the performance of the baseline models, the weighted ensemble, and the range of performances of the individual models in the ensemble. As well, we indicated the performance of the ensemble after removing models containing \( iBd_h \) and the performance of the ensemble after removing models containing \( iBd_h \) and/or \( iBd_d \). This was done to show the performance of the ensemble at forecast distances greater than one hour (but less than one day) and greater than one day, respectively.

6.3.2 Models for one location applied to the other

As mentioned, van Belle et al. (2007) found that models for one location did not predict well at a different location; however, the comparisons made in that study considered locations quite removed from one another (i.e. predictions for the Netherlands were made using models developed for southern Germany, Denmark, and southern Sweden). As well, the predictor and response variables were not necessarily consistent between the sites. Migration intensity was measured according to different methods using different devices (e.g. infrared device, pencil-beam radar, military surveillance radar), and meteorological variables were obtained from different sources. We have made bird density measurements in a consistent manner using two radars that are very similar to one another and not so far removed from one another in space (\(~98\) km). As well, we have obtained meteorological predictor variables for each site from the same data set. Because our study differed in these aspects from the study of van Belle et al. (2007), it was informative to apply our baseline and ensemble models, calibrated using data from one radar site, toward the prediction of migration intensity at the other radar site. We indicated all permutations of model performance (i.e. baseline, individual models, ensemble, ensemble without models containing \( iBd_h \), and ensemble without models containing \( iBd_h \) and/or \( iBd_d \)) for models calibrated to one location and used to predict intensity at the alternative location.

6.4 Results

6.4.1 Migratory dynamics and predictor variables

Time series of migration intensity, temperature \((T)\), tailwind assistance according to \( EQ^{\text{Tailwind}} \), and precipitation \((R)\) are shown in Figure 6.2 for De Bilt and Den Helder. Migration was generally more intense at De Bilt than Den Helder, but at both locations intense migration was the exception rather than the norm. Temperatures were slightly cooler in spring compared to autumn but were rarely below freezing at either site. Winds were more frequently
supportive in spring than autumn. Note also that tailwind support oscillated strongly at Den Helder in autumn.

Circular frequency distributions of the measured tracks and calculated headings from De Bilt and Den Helder are shown in Figure 6.3. The circular mean of these tracks and headings, also indicated in Figure 6.3, was used to classify winds as either supportive or unsupportive according to $E_{\text{Tailwind}}$. This circular mean is quite representative in most cases; however, in Den Helder during spring diurnal and autumn nocturnal migration there are two rather distinct groups of track directions (apparent from the bimodal distributions in Figure 6.3) with the circular mean falling between the two. This may somewhat confound the calculation of $W_{\text{acc}}$ and $RW_{\text{acc}}$ in these cases. Recall that tailwind support oscillated strongly at Den Helder in autumn (see Figure 6.2). This was probably not due to rapid changes in wind condition but to the different preferred directions used to calculate tailwind support between day and night (see Figure 6.3).

### 6.4.2 Baseline intensity

The functional relationship of the baseline intensity variable ($s_{\text{Time}}$) to migration intensity is shown in Figures 6.4 and 6.5 for De Bilt and Den Helder, respectively. $s_{\text{Time}}$ was included in all models by default. The performance of all models (including models containing only the variable $s_{\text{Time}}$) are indicated in Figures 6.7 and 6.8 and plots of measured against predicted migration intensity are shown in Figures 6.9 and 6.10 for De Bilt and Den Helder, respectively. Although models were calibrated independently for diurnal/nocturnal and spring/autumn migration, the baseline intensity rather smoothly transitioned from one model to the next (apparent from the general seamlessness of the contours between the plots in Figures 6.4 and 6.5). Nocturnal migration in spring was generally most intense during April at both locations, and nocturnal migration in autumn was most intense around mid-October in De Bilt but closer to the first of November at Den Helder. At both locations, nocturnal migration was more concentrated to a particular part of the season during autumn compared to spring. The most intense diurnal migration tended to occur near sunrise; however, this may very well reflect the end of nocturnal migration rather than the beginning of diurnal migration.

### 6.4.3 Autocorrelation in migration intensity

Plots indicating autocorrelation in $iBd$ measurements to a lag of one week are shown in Figure 6.6 for both radar sites. The $iBd$ measurements were positively autocorrelated at lags of one hour and one day in all cases, so our
Den Helder

Results

- Rainfall (R [mm])
- Tailwind (m/s)
- Temperature (T [°C])
- Birds per km²

Data for 2008 and 2009
Figure 6.2 (previous two pages): Time series plots for De Bilt (first) and Den Helder (second) indicating (from top to bottom) precipitation ($R$), tailwind assistance, temperature ($T$), and migration intensity. For plots of tailwind assistance, a horizontal gray line at -7 indicates the transition between winds considered supportive and unsupportive. In all other plots, this horizontal gray line indicates a value of zero.

Figure 6.3: Circular frequency distributions of measured track directions (dashed lines) and calculated headings (solid lines) that were associated with airspeeds determined to be between 7 and 25 ms$^{-1}$. An arrow in each plot indicates the circular mean of the tracks and headings, which was used to calculate $E_{\text{Tailwind}}^Q$ and determine whether wind conditions were supportive or unsupportive. A compass in the center of the figure indicates direction.
Figure 6.4: Plots illustrating the functional form (i.e. partial contribution) of the baseline intensity variable ($sTime$) in models of migration intensity in De Bilt. Spring conditions are shown in the left column, and autumn conditions are shown on the right. Diurnal migration is shown in the top row and nocturnal migration in the bottom. $sTime$ is represented as a smoothed interaction between the day of the year ($DOY$) along the x-axis and the fraction of the day ($P_{day}$) or night ($P_{night}$) relative to sunrise and sunset along the y-axis. Colors and contours indicate predicted migration intensity on the scale of the linear predictor. Predictions on the scale of the response variable (i.e. $iBd$) can be obtained by summing the partial contributions of all predictor variables in a model (in this case only $sTime$) and then applying the exponential function. Semi-transparent circles indicate $iBd$ measurements and their size reflects the value of $iBd$. 
Figure 6.5: Plots illustrating the functional form (i.e. partial contribution) of the baseline intensity variable \textit{sTime} in models of migration intensity in Den Helder. Spring conditions are shown in the left column, and autumn conditions are shown on the right. Diurnal migration is shown in the top row and nocturnal migration in the bottom. \textit{sTime} is represented as a smoothed interaction between the day of the year (DOY) along the x-axis and the fraction of the day ($P_{\text{day}}$) or night ($P_{\text{night}}$) relative to sunrise and sunset along the y-axis. Colors and contours indicate predicted migration intensity on the scale of the linear predictor. Predictions on the scale of the response variable (i.e. $iBd$) can be obtained by summing the partial contributions of all predictor variables in a model (in this case only \textit{sTime}) and then applying the exponential function. Semi-transparent circles indicate $iBd$ measurements and their size reflects the value of $iBd$. 
variables meant to capture autocorrelation (i.e. \(iBd_h\) and \(iBd_d\)) should be useful predictors of future migration intensity. In spring, migration intensity oscillated between being autocorrelated and not being autocorrelated through the one-day lag; whereas in autumn, migration intensity was positively autocorrelated (to greater and lesser degree) throughout the one-day lag. At De Bilt in fact, there was a peak in autocorrelation at a lag of 12-hours suggesting that migration intensity 12-hours previous was just as indicative of current migration intensity as migration intensity 24-hours previous. Thus, when migration was intense in autumn, it was intense during both the day and the night. In spring, however, diurnal and nocturnal migration occurred independently. The performance of all models (including models containing the variables \(iBd_h\) and \(iBd_d\)) are indicated in Figures 6.7 and 6.8 and plots of measured against predicted migration intensity are shown in Figures 6.9 and 6.10 for De Bilt and Den Helder, respectively.

6.4.4 Model performance

In Figure 6.7 the performance of our benchmark and ensemble models for De Bilt are shown alongside the performance of models applied to De Bilt that were calibrated for Den Helder. Similarly in Figure 6.8, the performance of our benchmark and ensemble models for Den Helder are shown alongside the performance of models applied to Den Helder that were calibrated for De Bilt. In both cases, we have indicated the range of performances of the individual models making up the ensemble and the weighted mean performance of the ensemble as a whole. The ensemble system was composed in total of 3473 unique models; however, 1093 models contained the variable \(iBd_h\) and were therefore only available to make forecasts one hour in advance. An additional 794 models contained \(iBd_d\) and were only available to make forecasts one day in advance. Therefore, we have also indicated the weighted mean performance of the ensemble at forecast distances greater than one hour (but less than one day) and at forecast distances greater than one day. The ensembles (at all forecast distances) generally performed better than the benchmarks composed of either \(sTime\) alone or \(sTime + iBd_d\), and these two benchmark models performed similarly to one another. The best model in the ensemble was generally the benchmark model composed of \(sTime + iBd_h\); however, in some cases other individual models in the ensemble performed slightly better (e.g. during autumn nocturnal migration at De Bilt, see Figure 6.7). The benchmark model composed of \(sTime + iBd_d + iBd_h\) also performed quite well, but it seemed that including \(iBd_d\) did little, if anything, to improve the performance of the \(sTime + iBd_h\) model. Note that for nocturnal migration in spring at De Bilt (see Figure 6.7), the weighted mean performances of the
CHAPTER 6. PREDICTING INTENSITY

Spring

Autocorrelation

De Bilt

Autumn

Den Helder

Spring

Autumn
6.4. RESULTS

Figure 6.6 (previous page): Plots indicating autocorrelation in $iBd$ measurements to a lag of one week at De Bilt (top two) and Den Helder (bottom two). Spring and autumn are shown separately for each radar. Dashed gray horizontal lines indicate the 99% confidence interval.

ensemble at all forecast distances were not visible. In this case, there were a few models that performed extremely poorly, which drove the weighted average performances beyond the range of values we showed in the plot. See the Discussion in Section 6.5 for more details.

Models calibrated to data from the alternative radar site performed worse than models calibrated for the actual site; however, in several cases these “alternative-site” models performed rather well. At De Bilt the alternative-site models performed only slightly worse than the actual-site models, but at Den Helder the alternative-site models were quite a bit worse than the actual-site models.

The magnitude of the MAD values appeared to show a strong dependence on the range of measured values, so absolute comparisons of the MAD values between sites and times should be made with caution. For example, one should not necessarily make the conclusion from Figures 6.7 and 6.8 that models of diurnal migration were better than models of autumnal migration even though MAD values were lower during the day compared to the night, because measured bird densities were also much smaller during the day compared to the night.

In Figures 6.9 and 6.10 we have plotted measured against predicted bird densities for De Bilt and Den Helder, respectively. In these plots, we show the predictions of two benchmark models, the full ensemble, the ensemble at a forecast distance greater than one hour (but less than one day), and the ensemble at a forecast distance greater than one day. All model permutations had a better fit to the data than the benchmark model containing the baseline intensity (i.e. $sTime$) only. In all cases, these models predicted the low intensity migration quite well but had difficulty predicting the few very intense migration events. Specifically, the models tended to underpredict the most intense migration events, and this occurred particularly when models did not contain $iBd_h$. 
CHAPTER 6. PREDICTING INTENSITY

Figure 6.7: The performance (i.e. mean absolute deviation or MAD) of models of hourly migration intensity for De Bilt are shown. The performance of these “actual-site” models was determined by 50-times repeated random sampling cross-validation. The performance of models calibrated for Den Helder and then used to predict at De Bilt (i.e. “alternative-site” models) is also shown. In both cases, the units of MAD are birds/km$^3$. White box plots and symbols indicate the performance of the actual-site models, gray box plots and symbols indicate the performance of the alternative-site models, and a dashed line separates the two. The ranges of performance of the individual models in an ensemble are shown as box plots distinguishing the median, inter-quartile range, one and a half-times the inter-quartile range beyond the quartiles, and outliers. Superimposed atop these box plots are the weighted mean performances of the full ensemble, the ensemble at forecast distances greater than one hour, and the ensemble at forecast distances greater than one day. Along the bottom of each box plot are the individual performances of the four benchmark models. A legend indicates the model that is represented by each symbol.
Figure 6.8: The performance (i.e. mean absolute deviation or $MAD$) of models of hourly migration intensity for Den Helder are shown. The performance of these “actual-site” models was determined by 50-times repeated random sampling cross-validation. The performance of models calibrated for De Bilt and then used to predict at Den Helder (i.e. “alternative-site” models) is also shown. In both cases, the units of $MAD$ are $\text{birds/km}^3$. White box plots and symbols indicate the performance of the actual-site models, gray box plots and symbols indicate the performance of the alternative-site models, and a dashed line separates the two. The ranges of performance of the individual models in an ensemble are shown as box plots distinguishing the median, inter-quartile range, one and a half-times the inter-quartile range beyond the quartiles, and outliers. Superimposed atop these box plots are the weighted mean performances of the full ensemble, the ensemble at forecast distances greater than one hour, and the ensemble at forecast distances greater than one day. Along the bottom of each box plot are the individual performances of the four benchmark models. A legend indicates the model that is represented by each symbol.
### De Bilt

<table>
<thead>
<tr>
<th></th>
<th>Spring</th>
<th>Autumn</th>
</tr>
</thead>
<tbody>
<tr>
<td>sTime</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sTime + IB&lt;sub&gt;d&lt;/sub&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E&lt;sub&gt;n&lt;/sub&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E&lt;sub&gt;n&lt;/sub&gt; &gt; 1 hr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E&lt;sub&gt;n&lt;/sub&gt; &gt; 1 day</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 6.9 (previous page): Scatter plots of predicted and measured bird densities for De Bilt are shown. Measured bird densities are always indicated along the x-axis and predicted densities along the y-axis, both in units of birds/km$^3$. Points are semi-transparent, so darkness indicates point density. A dashed diagonal line indicates a theoretical perfect positive relationship. Plots in the same column of the figure are from the same season and time (i.e. diurnal or nocturnal) and their axes are scaled similarly. Plots in the same row of the figure indicate a particular model. From top to bottom, the models are the benchmark model containing the baseline variable $sTime$, the benchmark model composed of $sTime$ and $iBd_h$, the full ensemble, the ensemble at a forecast distance greater than one hour, and the ensemble at a forecast distance greater than one day.

6.5 Discussion

6.5.1 Model development

General considerations

A model containing only the baseline intensity variable ($sTime$) was able to explain a large amount of variability in $iBd$ and performed better than some of the other models in the ensemble that contained more variables. If weather data and measurements of bird density were unavailable for some reason, the $sTime$ model could be used to give a general indication of the intensity of migration based only on the day of the year and the time of the day. Models that incorporated weather conditions generally performed better than the $sTime$ model, however, and also performed better than the benchmark model containing $sTime + iBd_d$, suggesting that migration intensity on a given day was better predicted by forecasted weather conditions than by migration intensity measured the previous day. Nonetheless, models that incorporated $iBd_h$ performed better still. This is perhaps to be expected, since $iBd_h$ reflects explicitly the migratory decisions birds in the area have already made. The relationships we uncover between migration intensity and environmental variables are informative, but we may never be able to account for all of the processes that have influence on migratory decisions at a given time and location. Currently, for instance, we cannot account for environmental conditions encountered earlier in the migratory journey that may have influenced current migratory decisions.

Clearly, using previous measurements of migration intensity can improve predictive models of future migration intensity; however, very recent measurements (i.e. $iBd_h$) improved predictions much more than older measurements (i.e. $iBd_d$). Regardless, an advantage of the use of weather radar (as opposed to military radar or dedicated bird-tracking radar, for example) is that data from weather radar are needed and valuable in many contexts, so a great deal
6.5. DISCUSSION

Figure 6.10 (previous page): Scatter plots of predicted and measured bird densities for Den Helder are shown. Measured bird densities are always indicated along the x-axis and predicted densities along the y-axis, both in units of $\text{birds/km}^3$. Points are semi-transparent, so darkness indicates point density. A dashed diagonal line indicates a theoretical perfect positive relationship. Plots in the same column of the figure are from the same season and time (i.e. diurnal or nocturnal) and their axes are scaled similarly. Plots in the same row of the figure indicate a particular model. From top to bottom, the models are the benchmark model containing the baseline variable $s\text{Time}$, the benchmark model composed of $s\text{Time}$ and $i\text{Bd}_h$, the full ensemble, the ensemble at a forecast distance greater than one hour, and the ensemble at a forecast distance greater than one day.

...of effort is put toward having these radars continuously operational. Thus, we can be relatively certain that previous measurements of migration intensity will be available for input into these models. Unfortunately, $i\text{Bd}_h$ is only available for forecast distances of one hour, which limits the operational applicability of models in the ensemble that contain this variable. Perhaps the two-hours-previous $i\text{Bd}$ measurement, which would be available for forecast distances of two hours, would also improve model performance. This seems plausible, since the autocorrelation plot in Figure 6.6 suggests a strong correlation still at a lag of two-hours.

...In accordance with the findings of van Belle et al. (2007), models in our study calibrated for a particular location performed better at that location than models applied to that location that were calibrated elsewhere. The distance between our two radar sites was not that great (~98 km), but the different positions of these two sites (perhaps relative to the coast) seemed to result in different migratory dynamics (see e.g. Figures 6.2 and 6.2 as well as Figure 6.3). Nonetheless, the performance in this study of models used in one location that were calibrated for an alternative location is encouraging, as in many cases these models still performed rather well (particularly when the variable $i\text{Bd}_h$ was included in the models). Naturally, areas covered by a weather radar should have ensemble models calibrated explicitly for their location; however, there are (and will likely continue to be) spatial gaps in radar coverage within which models cannot be explicitly calibrated. Our results suggest that models from nearby locations may be useful in filling these spatial gaps. Predictions for these gaps could be made using local environmental conditions as input into an ensemble prediction system calibrated for a nearby location. Alternatively, the predictions from models for several nearby locations could be spatially interpolated to fill these gaps. In either case, the more sensors with individual predictive models that are available, the more accurate information we will have to fill these gaps. For instance,
a large network of sensors will be useful in determining the spatial range in which measurements (and similarly predictions) are valid by allowing for the calculation of spatial autocorrelation. As well, such a network would allow for the use of modern spatial interpolation techniques such as kriging (cf. Hengl et al., 2007). Kriging would not only allow us to create a contiguous predictive surface over the spatial range of available sensors, potentially incorporating underlying landscape types as explanatory variables, it would enable the identification of particular geographic areas where the interpolation was less reliable. These aspects of kriging may be particularly beneficial since the distances over which measurements (and therefore predictions) are valid likely depend on particular features in an area such as the sizes and positions of large water bodies and/or mountains (cf. Åkesson et al., 1996; Bruderer and Liechti, 1998; Fortin et al., 1999). A further possibility to consider in filling gaps between sensors may be to use simulation models such as the FLAT model introduced in Chapter 3 to propagate birds measured in one location through space to other locations. When birds are observed by radar leaving the southern tip of Norway, for example, the FLAT model could simulate the birds’ continued movement through space to indicate when and where they will likely arrive on Great Britain or the European mainland. This approach has shown promise in previous research (Shamoun-Baranes and van Gasteren, 2011).

Our analyses were conducted using two autumn and two spring migration seasons. van Belle et al. (2007) determined that the longer the time series used to calibrate predictive models of bird migration, the more robust and accurate are the resulting models. The time series used in this study likely represent the minimum amount of data upon which the calibration of models is feasible. Models will certainly be more robust when calibrated in areas with longer time series of data. Furthermore, as more data become available for a particular location, the ensemble modeling system can (and likely should) be re-calibrated incorporating all available data. The model-development framework we have outlined can be used to ensure that the recalibration of these models is done efficiently and consistently.

Unique considerations

The model development procedure we have described is flexible enough to incorporate more predictor variables. Therefore, when developing models of migration intensity, predictor variables other than the ones mentioned here may be incorporated that reflect location-specific relationships between environmental conditions and migration intensity uncovered in previous research in the area. Regardless of whether more variables are included or not, how-
ever, issues are likely to arise during model development that are unique to the time period, location, and/or data set considered. These issues may require unique considerations or adjustments to the general modeling procedure outlined thus far. In this section, we discuss some of the issues that arose as we developed models for De Bilt and Den Helder and suggest potential approaches for dealing with these issues.

For nocturnal migration in spring at De Bilt, there were a few models that performed exceptionally poorly. While the median $MAD$ value of the models in the ensemble was 1.55, a few models had $MAD$ values in the hundreds, and the worst of these models had an $MAD$ value of 2703. These poorly performing models resulted in the weighted average performance of the entire ensemble being 2.44, therefore just outside the viewing area of the plot in Figure 6.7. GAMs can produce biased estimates near the edges of a domain or “edge effects” that are associated with higher-dimensional smoothing (see Webster et al. (2006), citing Hastie and Tibshirani (1990), which can result in large errors when predictions are made from the edges of these domains. There are several options for dealing with a subset of the ensemble that performs much worse than the rest. One option is to allow the models to remain in the ensemble. The models are weighted in the ensemble according to their performance, so poorly performing models carry less weight in the overall predictions. However, when some models perform very poorly, models that perform moderately poorly will perhaps carry more weight in the ensemble than is desirable. Another straight-forward option is to remove all models from the ensemble with performances that are considered to be outliers from the performance of the rest of the models in the ensemble. In our case, for example, we could remove all models from the ensemble with $MAD$ values that were larger than one-and-a-half times the inter-quartile range beyond the upper-quartile, which is an approach that could be automated and kept consistent. For nocturnal migration in spring at De Bilt, applying this procedure would have resulted in the removal of 262 models with $MAD$ values greater than 2.32. Still another option is to identify the variable that is causing the problems and limit the amount of smoothing that the GAM fitting procedure is allowed to apply to the variable. We found, for instance, that all of the models exhibiting this abysmal performance contained the variable $prcp$. After identifying the troublesome variable and restricting the amount of smoothing allowed, the performance of all models containing that variable must be recalculated in order to determine their new weight in the ensemble. This approach would likely require manual intervention into the model-development procedure.

Variables that need to be set per site can be problematic and require sufficient data and perhaps also knowledge of the migratory dynamics in an
area. Therefore, a potential limitation on the exportability of the model-development procedure we have outlined is in the calculation of the accumulation variables, particularly $W_{acc}$ and $RW_{acc}$. While the methods we have applied to calculate these accumulations is exportable, the settings of particular parameters may be more or less applicable depending on the migratory dynamics of a particular time and location. In our analysis, the preferred direction of migration we determined was representative of the distributions of tracks and headings in most cases; however, at Den Helder during spring diurnal and autumn nocturnal migration there appeared to be two distinct groups or ‘cohorts’ exhibiting different track directions (see Figure 6.3). The circular mean of the tracks and headings at these times fell between the two cohorts, and, therefore, the preferred migratory direction assumed in these cases was not explicitly representative of either cohort. While the two cohorts were not separated by 180° such that all supportive winds for one cohort were prohibitive for the other and vice versa, some wind conditions supportive of one cohort were probably prohibitive for the other and vice versa. In locations where this is an issue, it may be beneficial to calculate an accumulation due to wind for each cohort individually, each based on a preferred direction of migration calculated according to the circular mean of the tracks and headings of a particular cohort. Another potential issue is the threshold at which winds are considered supportive or unsupportive, which was set to $EQ^{Tailwind} = -7$ in our analyses. We observed a step-change in $iBd$ values when they were plotted against wind support according to $EQ^{Tailwind}$ (not shown), and this step-change occurred near a value of -7, but this value may not be representative in other locations. It is therefore advisable to plot $iBd$ against wind support for a location to determine an appropriate threshold. Regardless, accumulation due to successive days of bad weather remains difficult to capture in models and is likely influenced by conditions encountered earlier in the migratory journey and ‘upstream’ of the measurement location. An integrated network of sensors should be beneficial in this context as information from upstream radars can be incorporated into downstream models.

6.5.2 Flight safety

A primary use of the models developed in these analyses was to forecast migration intensity for flight safety. In the models that resulted from this procedure, small and medium intensity migration (which comprise the majority of measurements) are predicted quite well (see Figures 6.9 and 6.10); however, the very infrequent instances of intense migration are not well-predicted. Specifically in the context of flight safety, where the peak migration events are quite important, it may be beneficial to add additional weight during model calibra-
tion to observations exhibiting more intense migration, and we explored the possibility of doing so in our analyses. We found that the peaks were better represented in the resulting models, since the weighting procedure resulted in response variables that were parameterized to fit the peaks of the calibration data set very well; however, it is questionable whether or not these peaks will occur under similar conditions in subsequent years. For example, in order to compensate for the few intense (but very influential, due to the weighting) bird density measurements, the functional form of the baseline intensity variable \((sTime\); shown in Figures 6.4 and 6.5) becomes very distorted. It is unlikely that migration will be intense on precisely the same day and time the following year because (among other things) the atmospheric conditions will likely be quite different. The functional forms of the other variables in the models are likely to be similarly distorted in ways that are not representative of their actual influence on migration intensity. This method of weighting more intense migration events in the calibration of models may become more feasible as more data become available and intense migration can be observed to occur (and not occur) through a more representative range of the domains of each predictor variable.

Until more data are available, a temporary solution may be to merge the weighted and unweighted models into a single ensemble system. The models developed without weighting perform well and accurately reflect the majority of the data. The models developed with weighting applied in proportion to the intensity of migration better capture the peak migration events. Thus, an optimal solution may be a hybrid of the weighted and unweighted model ensembles. A potential method to merge the two ensembles would be to make a prediction by each individually, determine the average of those predictions, and use this average to determine how much influence the weighted and unweighted ensembles should be given in the final prediction. The lower the average, the more influence given to the unweighted models; the higher the average, the more influence given to the weighted models. Figure 6.11 illustrates this concept. The influence of the predictions from the weighted model ensemble can be calculated as

\[
\tan^{-1}\left(\frac{a \cdot x - a \cdot b}{\pi}\right) + 0.5, \quad (6.4)
\]

and the influence of the predictions from the unweighted model ensemble can be calculated as

\[
1 - \tan^{-1}\left(\frac{a \cdot x - a \cdot b}{\pi}\right) + 0.5 \quad (6.5)
\]

where \(x\) indicates the average of the predictions of the weighted and un-
Figure 6.11: This figure illustrates the concept of predictions being based on the influence of two ensemble systems: one with weighting applied in the calibration process to better represent very intense migration and one without weighting applied in the calibration process to better represent the smaller (and more frequently occurring) bird densities. When the average of the predictions of the two ensembles is small, the final prediction is based on the unweighted model ensemble. As the average prediction of the two ensemble systems increases, the final prediction is based more and more on the weighted model ensemble.

weighted ensembles, $a$ indicates how fast the transition between the two model-types occurs, and $b$ indicates the inflection point or the value of $x$ at which the influence of the weighted and unweighted ensembles is equal.

The appropriate amount by which to weigh measurements of bird density in the calibration of models may depend on the particulars of the data set. We found that weighing each $iBd$ measurement by $iBd^{1.25}$ produced decent results. The parameters $a$ and $b$, controlling the influence of the weighted and unweighted ensembles should be calibrated to produce the most accurate predictions for a given time and location.

6.6 Conclusion

With the model-development framework outlined in this chapter, models of migration intensity can be systematically developed for new locations. The models that result from this procedure can be used to forecast migration intensity up to several days in advance, which is generally the valid range of the numerical weather forecasts upon which most of the models in the ensemble depend. As well, our results suggest that the models of migration intensity can provide useful information on migratory dynamics in nearby locations, particularly for short-term forecasts. This is particularly useful for locations that are not covered by a weather radar and therefore have no measurements of
bird density with which to calibrate unique models. The study also shows that incorporating (particularly recent) measurements of bird density into forecast models improves their performance.

This study highlights the potential benefits of extracting bird density information from operational weather radar and provides a system to develop predictive models of bird density in new locations as data from these radars become available. Ultimately, the measurements of bird density from individual weather radars (and the models developed from them) should be integrated into unified large-scale monitoring and prediction systems that will dramatically improve flight safety (for both military and civil aviation) and likely revolutionize the field of ornithology.

Acknowledgments
We would like to thank the Royal Netherlands Meteorological Institute (KNMI) for providing data from the De Bilt and Den Helder weather radars and Toon Moene, Hidde Leijnse, Adriaan Dokter, and KNMI for providing and assisting with HIRLAM weather data. We also thank Hidde Leijnse, Hans van Gasteren, James McLaren, and Adriaan Dokter for invaluable discussion and contributions throughout. NCEP/DOE Reanalysis II data were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA [http://www.esrl.noaa.gov/psd/]. Our studies are facilitated by the NLeSC [http://www.esciencecenter.com/] and BiG Grid [http://www.biggrid.nl] infrastructures for e-Science and supported financially by the Ministry of Defense (Flysafe2).