Chapter 9
Validation Experiments

The variable rate model developed for MH370 was validated by analysing data from a collection of flights where the true aircraft location was known; we refer to these as validation flights. A total of six validation flights were used for testing. Data was available from a larger number of flights but the majority of these were in relatively short segments of less than three hours. There were only a few that maintained communications with the satellite Inmarsat-3F1 for longer periods and it was not thought productive to examine the prediction performance over time segments shorter than three hours. Of the six flights, four are previous flights of the accident aircraft, 9M-MRO, and the other two are flights of different aircraft that occurred at the same time as the accident flight. Three of the flights are relatively short and are between locations inside Asia, and the other three are flights from Asia to Europe.

The data available for the accident flight consists of mostly R1200 communication messages at approximately one hour intervals. In order to emulate the measurement information content, measurement data sets were formed by randomly sub-sampling R1200 communication messages from the validation flights. Ten different subsets were formed for each validation flight, resulting in a total of sixty validation measurement sets. Multiple sets were drawn from each flight to increase the statistical significance of the testing data set. They also serve to illustrate the sensitivity of the method to the precise measurement times and values. The measurement subsets were selected using a randomised process that aimed to achieve an average time between measurements of one hour. For the analysis we treat the measurement subsets as independent Monte Carlo trials. However there are several variables that are in common within the group of ten subsets of a single flight: the aircraft geometry is obviously the same for each subset since they are drawn from the same flight; the residual wind errors are the same; and the BFO is known to have a slowly varying bias, so there can be correlation in the BFO measurements from different subsets if those subsets choose measurements at similar times. Finally, some subsets may in fact randomly choose the same measurement as another subset.

In each validation flight, the true aircraft location was obtained from the Aircraft Communications Addressing and Reporting System (ACARS) data logs. Sections of the flight immediately after take-off and prior to landing were not included in the analysis since the aircraft dynamics are very different at these times and it is unlikely that sparse satellite messages would be sufficient to follow it. For the longer flights into Europe, the aircraft changed satellites part way through the flight so it was not possible to use the whole flight: these were truncated near the end of messaging via the Indian Ocean Region satellite. The filter was initialised using the true aircraft location, speed and control angle with a Gaussian random error. The standard deviation of the initialisation error was chosen to be the same as the prior for the accident flight, that is 0.4° in latitude and longitude, 1° in angle and Mach 0.03 in air speed. For every subset the posterior pdf at the final measurement was predicted ahead to a common time, corresponding to an exact ACARS reporting time. This predicted pdf is compared with the ACARS report.

This chapter first explains the particular characteristics of each flight and presents an example output pdf for one of the measurement sets. This output is subjectively compared with the ACARS truth. The statistical analysis is then presented using an objective performance measure over the sixty validation subsets. Table 9.1 lists the six validation flights used for the analysis and gives comments on some of the characteristics of each. The flights are ordered by time.

Table 9.1 Summary of validation flights

Flight path	Date	Duration (h:mm)	9M-MRO	Comments
Kuala Lumpur to Amsterdam	26 February	7:35	Yes	Eclipse
Mumbai to Kuala Lumpur	2 March	3:20	Yes	Short and almost straight with a gradual late veer, outlier BFO measurements
Kuala Lumpur to Beijing	6 March	4:25	Yes	Single climb, several S-turns
Beijing to Kuala Lumpur	7 March	4:55	Yes	Significant climbs, Mach changes and turns, contains anomalous BTO measurements
Kuala Lumpur to Amsterdam	7 March	7:50	No	Large S-turns
Kuala Lumpur to Frankfurt	7 March	7:03	No	Mid flight heading deviations, outlier BFO measurements

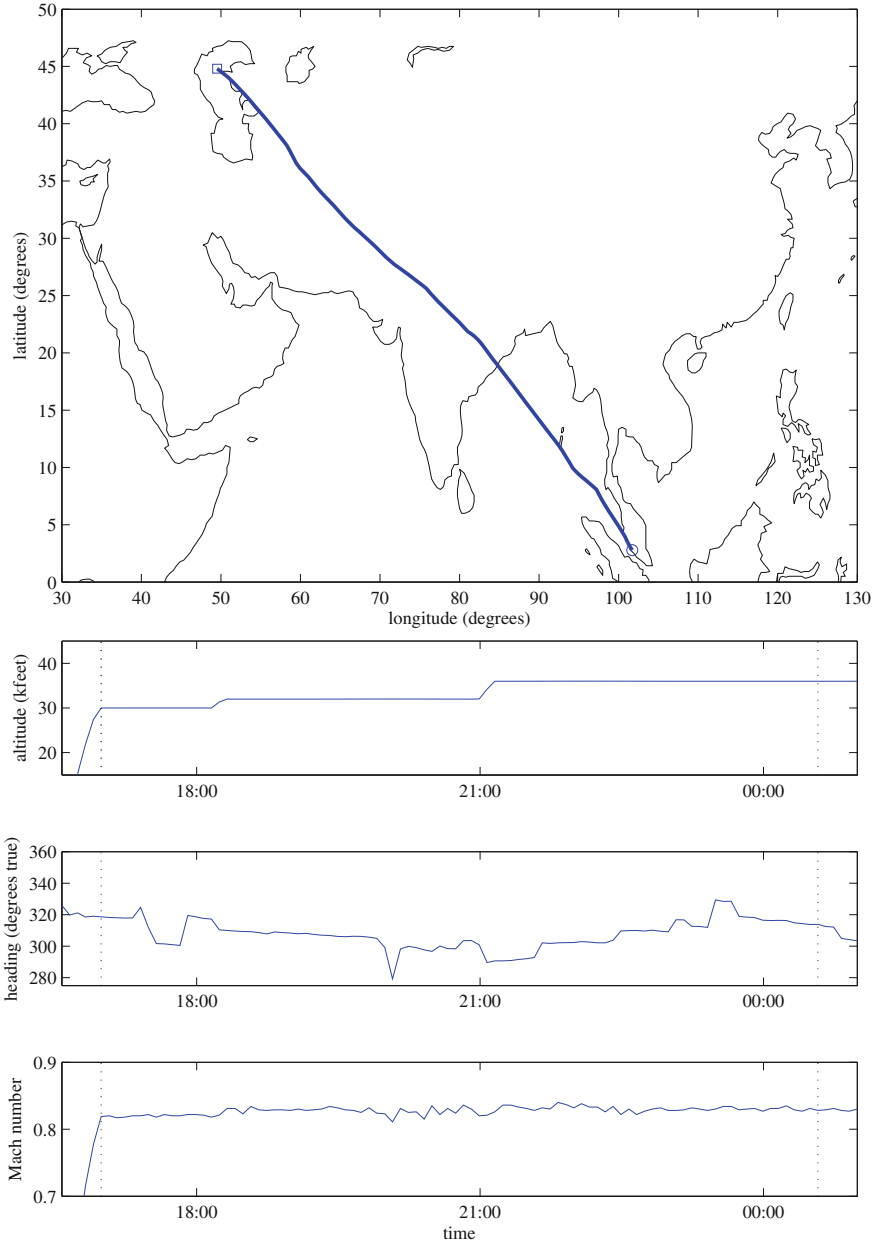## 9.1   9M-MRO 26 February 2014 Kuala Lumpur to Amsterdam

The first validation flight was from Kuala Lumpur to Amsterdam on 26 February 2014. This flight was around 7.5 h long but is relatively straight. Figure 9.1 summarises the features of the flight: the upper plot shows a geographic plan; the lower three plots show the aircraft altitude as a function of time, the aircraft heading as a function of time, and the aircraft Mach number as a function of time. Vertical dotted lines show the start and end of the time segment selected for the test. This flight contained an eclipse event so the validation also supports the Inmarsat eclipse correction [2].

Figure 9.2 shows the filtered pdf for the Kuala Lumpur to Amsterdam flight visualised using a three dimensional representation in Google Earth. The filter pdf is defined over a high dimensional space but for visualisation we examine the marginal position distribution in latitude and longitude. Because the BTO measurement error is relatively small the position distribution is centred on an arc of zero BTO error and has a narrow off-arc width. For the visualisation we marginalise the distribution onto the zero BTO error arc and encode the probability density for each point along the arc using altitude: points on the curve higher above the earth correspond to higher probability. A white curve on the map marks the ACARS reported aircraft location, a yellow marker denotes the location of the aircraft at each measurement time. The figure also shows a representative selection of the paths sampled by the filter. The selection shows the highest probability path arriving at each point around the arc: the colour of the path shows the marginal probability at that location on the arc (using a colour map similar to Fig. 5.7, i.e., blue is least likely, red is most likely).

There are a number of paths that end in significantly different locations to the truth. These occur because in this flight the aircraft travels in a direction that is almost horizontally radial from the satellite. While the aircraft moves towards the satellite its initial dynamics constrain the plausible paths but once it passes through the point of closest approach and begins to move away then it is possible to make turns that result in different near-radial paths. The support of these ambiguous paths is disjoint because of the finite number of samples: the true underlying pdf has support all the way around the arc. Without dynamic constraints the location of the peak of the pdf is simply a function of measurement noise.
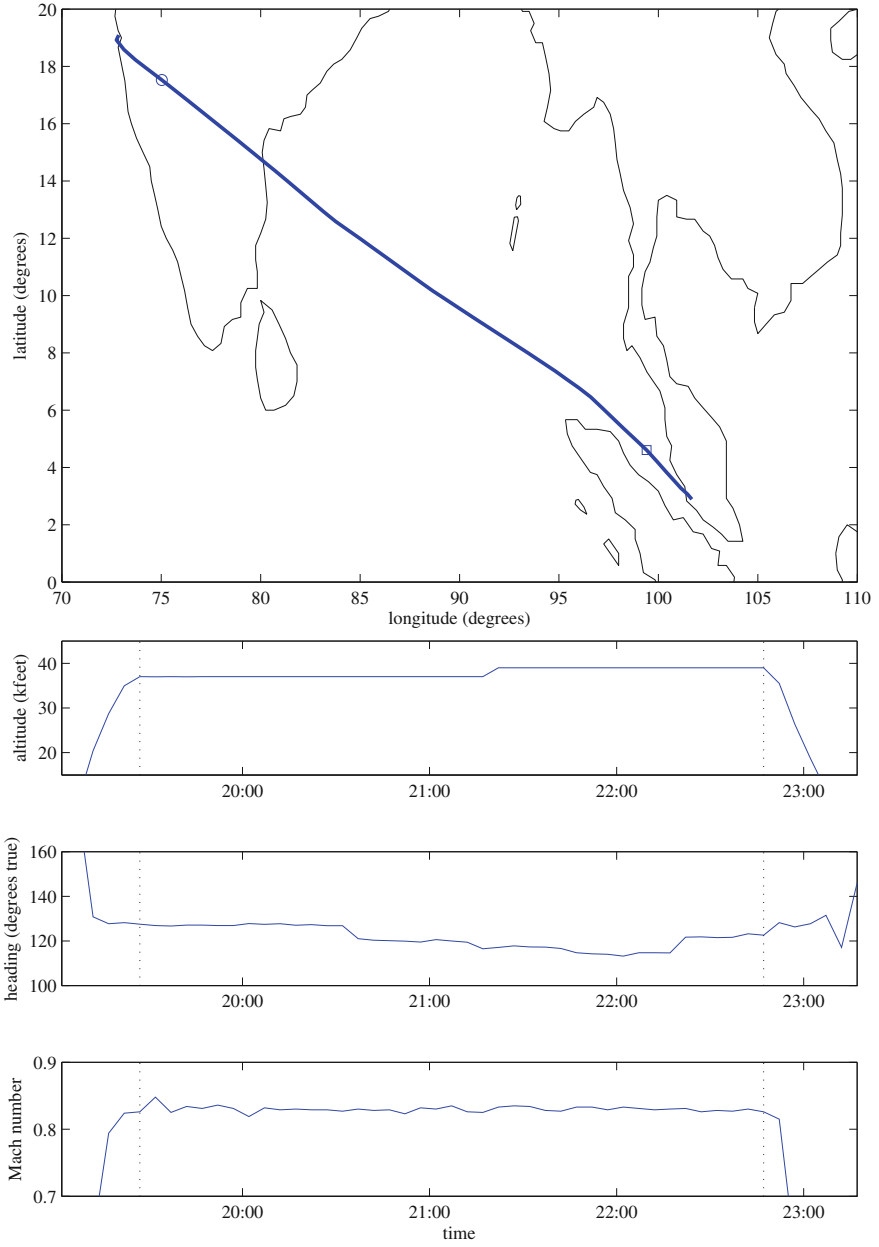
9.2   9M-MRO 2 March 2014 Mumbai to Kuala Lumpur

The flight from Mumbai to Kuala Lumpur is the shortest validation flight selected. Figure 9.3 summarises the features of the flight: there is a single minor altitude change and the Mach number remains relatively constant. The aircraft heading gradually reduces for most of the flight, turning the aircraft more to the North but a veer near the end turns it back to the South-East. The BFO measurements for this flight contain several outliers that are more than 30 Hz away from other measurements at similar times.

Fig. 9.1 Validation flight 26 February 2014 Kuala Lumpur to Amsterdam. *Vertical dotted lines* show the start and end times of the segment used for validation

Fig. 9.2   Validation flight 26 February 2014 Kuala Lumpur to Amsterdam

Fig. 9.3 Validation flight 2 March 2014 Mumbai to Kuala Lumpur. *Vertical dotted lines* show the start and end times of the segment used for validation

Figure 9.4 shows the pdf output from the filter; the true ACARS aircraft location is again under the main peak of the pdf. The pdf might appear to be relatively spread compared with some of the other flights, but the scale is much smaller in this case because the flight is short.

9.3   9M-MRO 6 March 2014 Kuala Lumpur to Beijing

This flight is the MH370 route from Kuala Lumpur to Beijing that was flown by the accident aircraft 9M-MRO on 6 March 2014, i.e., the day prior to the accident flight. Figure 9.5 summarises the features of the flight: the flight contained a single altitude change and several turns. Observe that there are several times where the heading changes for a short time before reverting back to the previous long-term value. These course corrections have the impact of translating the flight path and then returning to the previous ground velocity vector: in effect they are a kind of S-turn manoeuvre. If one or more of these corrections occurs between measurements then the most likely paths can be biased because there are no measurements to hint that the manoeuvres have occurred and the S-turn trajectory is less probable under the dynamics model than a constant angle path.

Figure 9.6 shows the pdf from the filter: the pdf is multi-modal with three main peaks that are somewhat blurred together. There was a heading change just before the last measurement and the lack of future data makes it impossible to resolve exactly what manoeuvre led to the change in range rate. One of the peaks of the pdf is clearly centred close to the true location.

9.4   9M-MRO 7 March 2014 Beijing to Kuala Lumpur

This flight is the MH371 route that was flown by the accident aircraft 9M-MRO on the morning of 7 March 2014 and is the return flight from Beijing back to Kuala Lumpur. Figure 9.7 summarises the features of the flight: there were three altitude changes and two main heading changes, the first of which was almost immediately after the start of the validation segment. This flight does not contain the S-turn manoeuvres that were present in the previous flight. In addition to the altitude changes the Mach number of the aircraft changed from 0.83 to 0.82. Each of these leads to a change in air speed. This flight contained several anomalous BTO measurements that were corrected using the empirical adjustment described in Chap. 5.

Figure 9.8 shows the pdf output from the filter; the true ACARS aircraft location is again under the main peak of the pdf. The peak is more spread because the altitude changes and Mach change modify the radial speed between the aircraft and the satellite. The resulting BFO measurements can also be explained by course changes: the aircraft could change speed or it could turn slightly. The BFO measurement is not informative enough to discriminate strongly between these and there is not enough subsequent data to see which is more consistent with BTO progression.
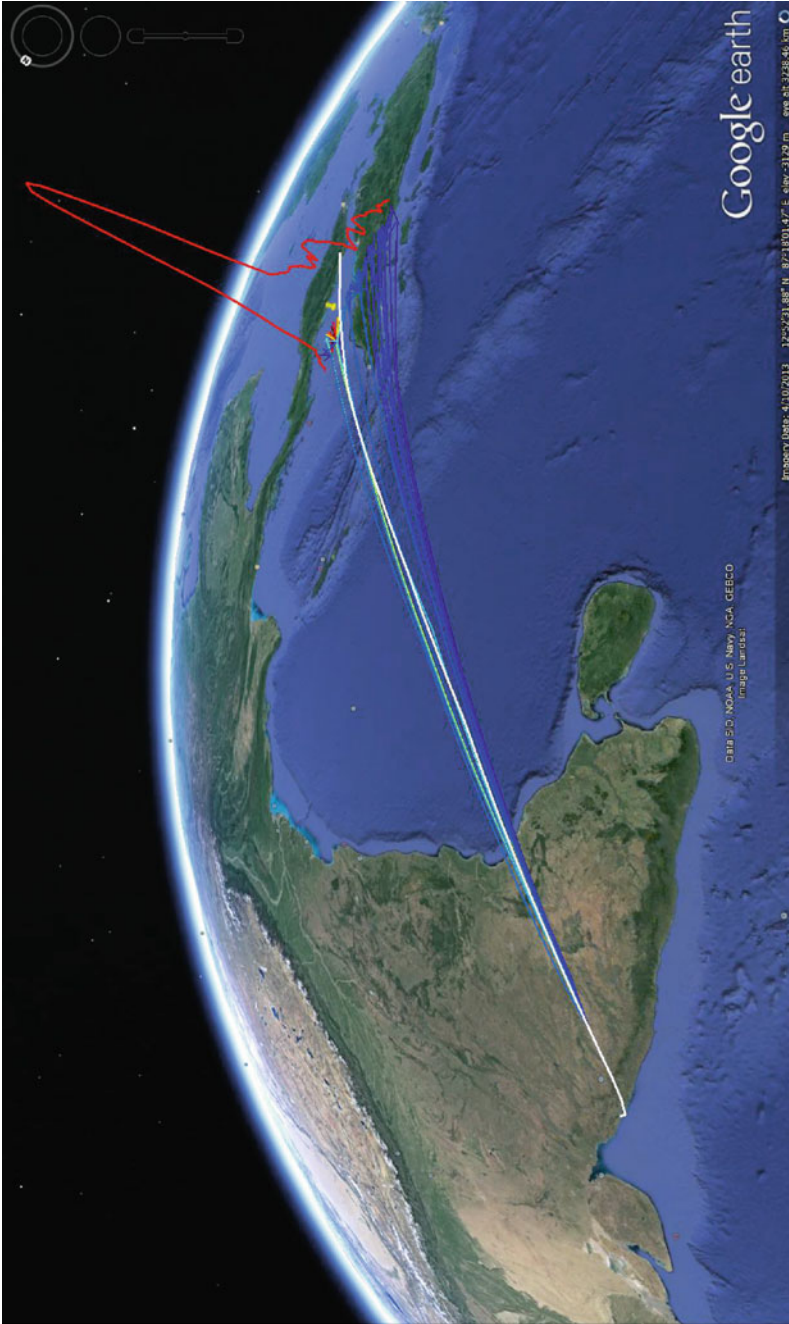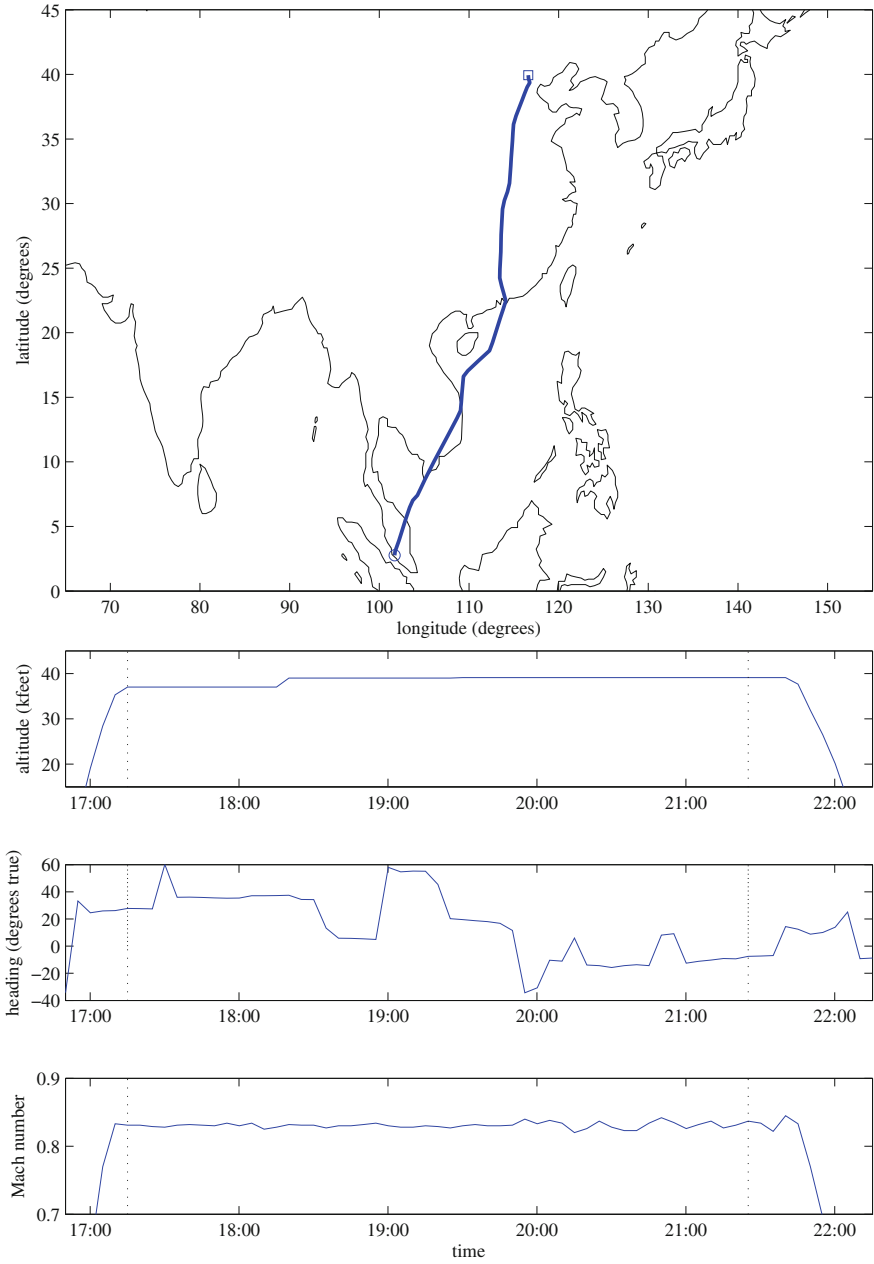
Fig. 9.4   Validation flight 2 March 2014 Mumbai to Kuala Lumpur

Fig. 9.5 Validation flight 6 March 2014 Kuala Lumpur to Beijing. *Vertical dotted lines* show the start and end times of the segment used for validation
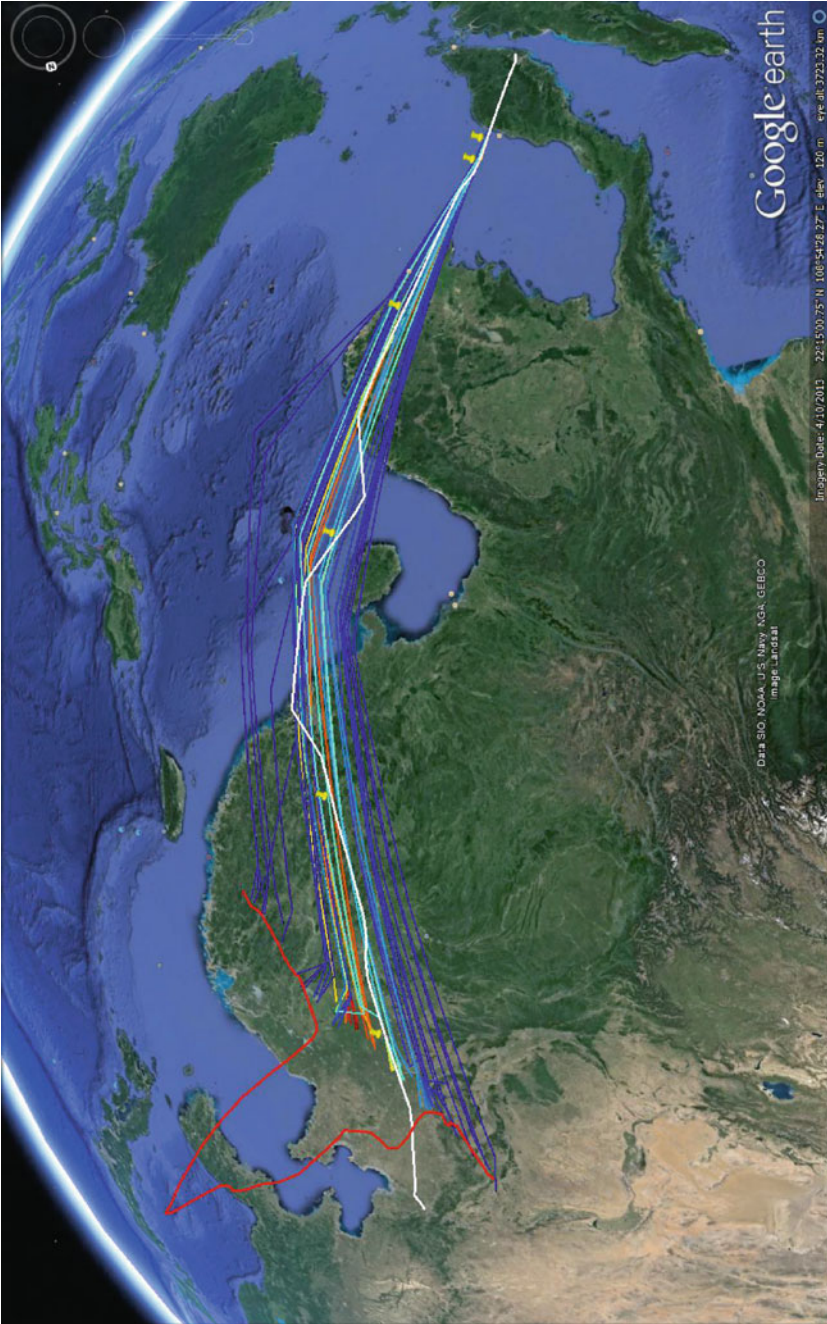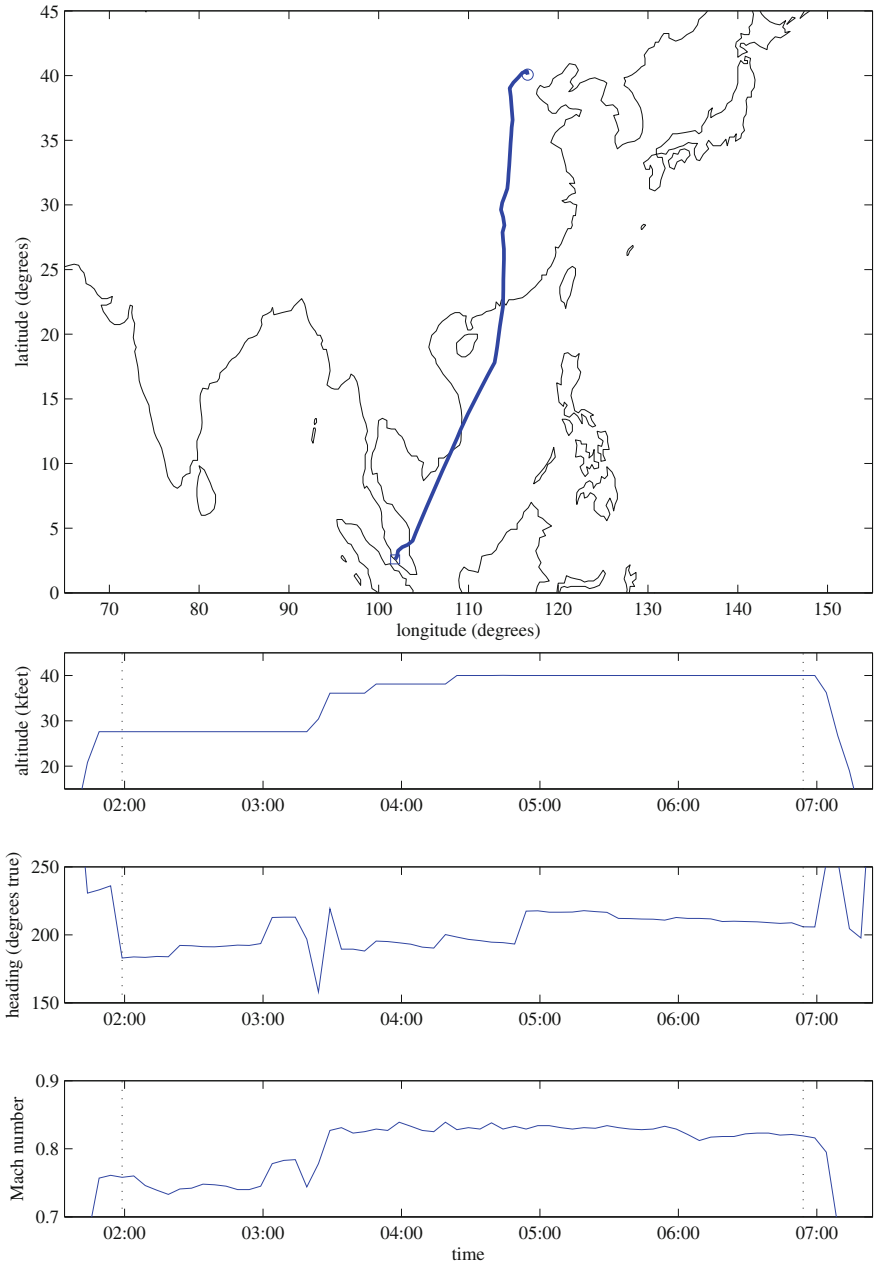
Fig. 9.6 Validation flight 6 March 2014 Kuala Lumpur to Beijing

Fig. 9.7 Validation flight 7 March 2014 Beijing to Kuala Lumpur. *Vertical dotted lines* show the start and end times of the segment used for validation
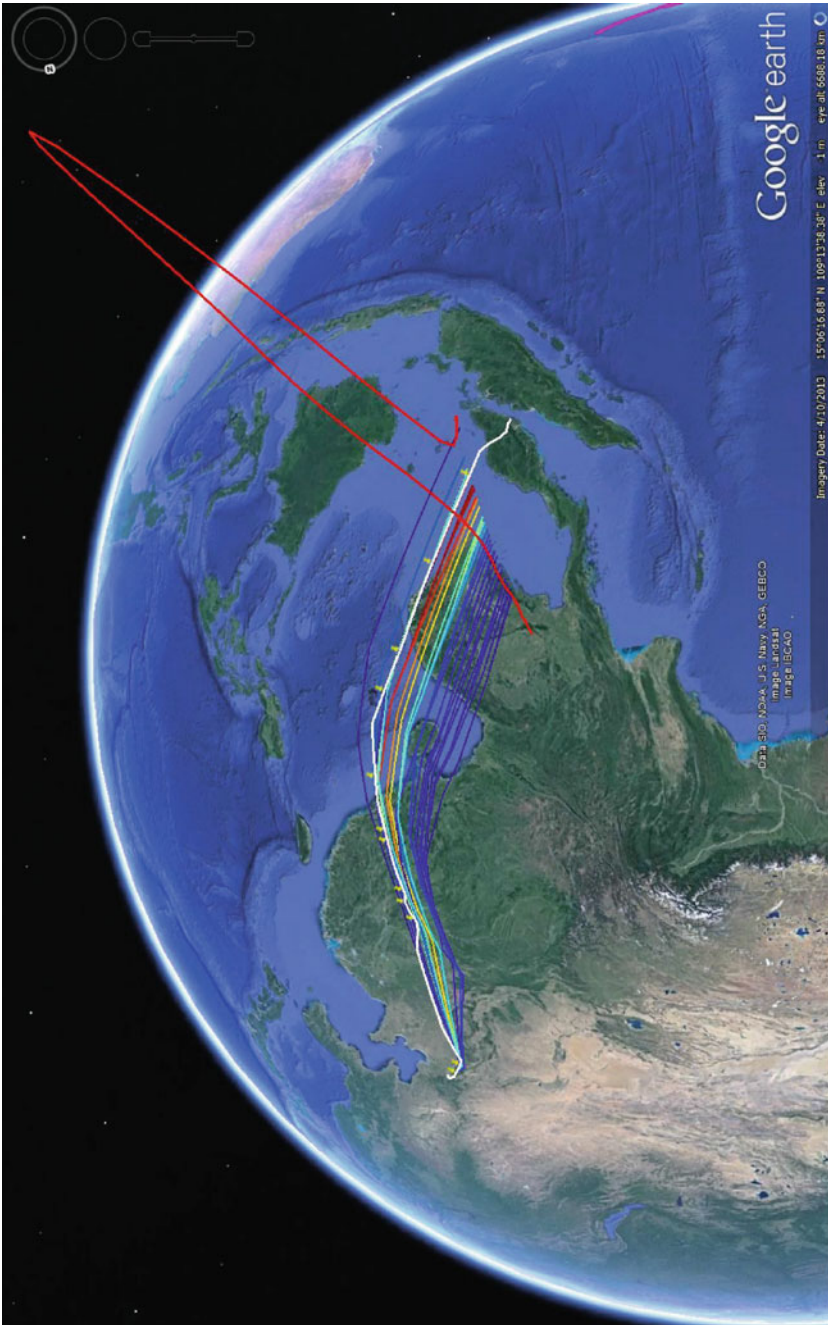
Fig. 9.8   Validation flight 7 March 2014 Beijing to Kuala Lumpur

9.5   7 March 2014 Kuala Lumpur to Amsterdam

This flight was from Kuala Lumpur to Amsterdam and is the same flight path as the first validation flight but with a different aircraft. Figure 9.9 summarises the features of the flight: the aircraft climbs with a sequence of vertical manoeuvres and there is a large S-turn manoeuvre near to the end of the analysed flight segment.

Figure 9.10 shows the pdf output from the filter. The true ACARS aircraft location is under the main peak of the pdf but in this case the true location is lower in the tails than in the other cases. The numerical results that follow in Sect. 9.7 show that this flight had the worst overall performance of the validation flights, although, as discussed in the next section, for each subset of measurements, the final location is within the region containing 85 % of the probability distribution, i.e., the highest posterior density (HPD) interval, discussed further in Sect. 9.7.2.
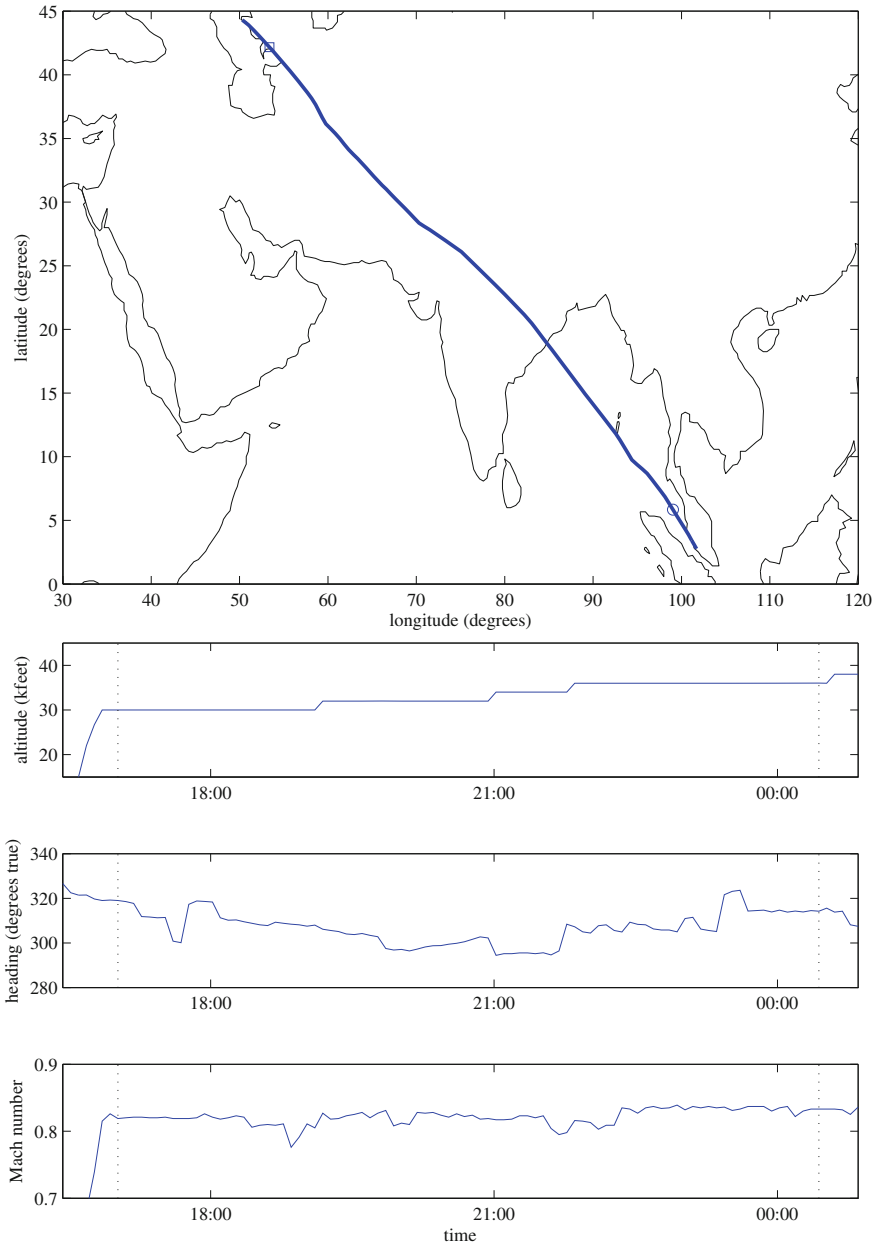
9.6   7 March 2014 Kuala Lumpur to Frankfurt

The final validation flight was from Kuala Lumpur to Frankfurt. Figure 9.11 summarises the features of the flight. It shows the full flight path, but the communications satellite changes part way through and the test section finishes where the box is marked on the map. No Mach information was available for this flight. There was a large heading deviation mid-flight, but the aircraft eventually reverted back to the earlier heading: this kind of compound manoeuvre is difficult for the filter to characterise. This flight also contained outlier BFO measurements.

Figure 9.12 shows the pdf output from the filter. The performance on this flight is quite similar to the Kuala Lumpur to Amsterdam flights. The filter has again identified ambiguous paths due to the relative geometry.

9.7   Quantitative Analysis

The examples above present a qualitative measure of performance but a more rigorous objective measure is required to provide a statistical assessment of the filter output. So far we have been satisfied that the true location has been in an area of reasonable support for the pdf, but is the spread of the pdf appropriate and is the mode of the distribution biased? Answers to questions such as these require a much larger ensemble of test data. However, it has not been feasible to collect the required test measurements for dozens of different flights. In order to increase our confidence in the performance for the relatively small set of flights that is available, multiple communication measurement sets were generated for each flight by randomly selecting individual R1200 messages from the communication logs of each flight.

Fig. 9.9 Validation flight 7 March 2014 Kuala Lumpur to Amsterdam. *Vertical dotted lines* show the start and end times of the segment used for validation

Fig. 9.10 Validation flight 7 March 2014 Kuala Lumpur to Amsterdam
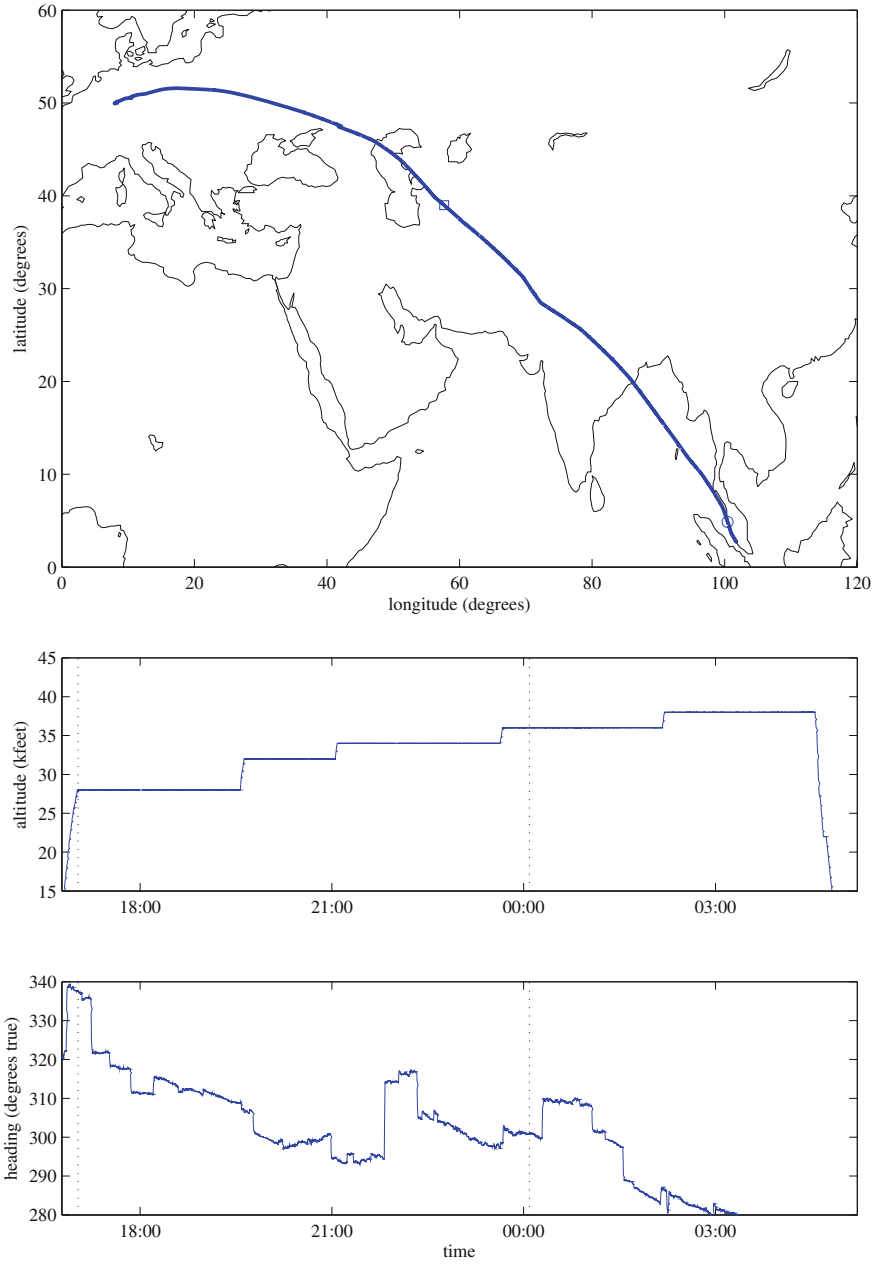
Fig. 9.11 Validation flight 7 March 2014 Kuala Lumpur to Frankfurt. *Vertical dotted lines* show the start and end times of the segment used for validation
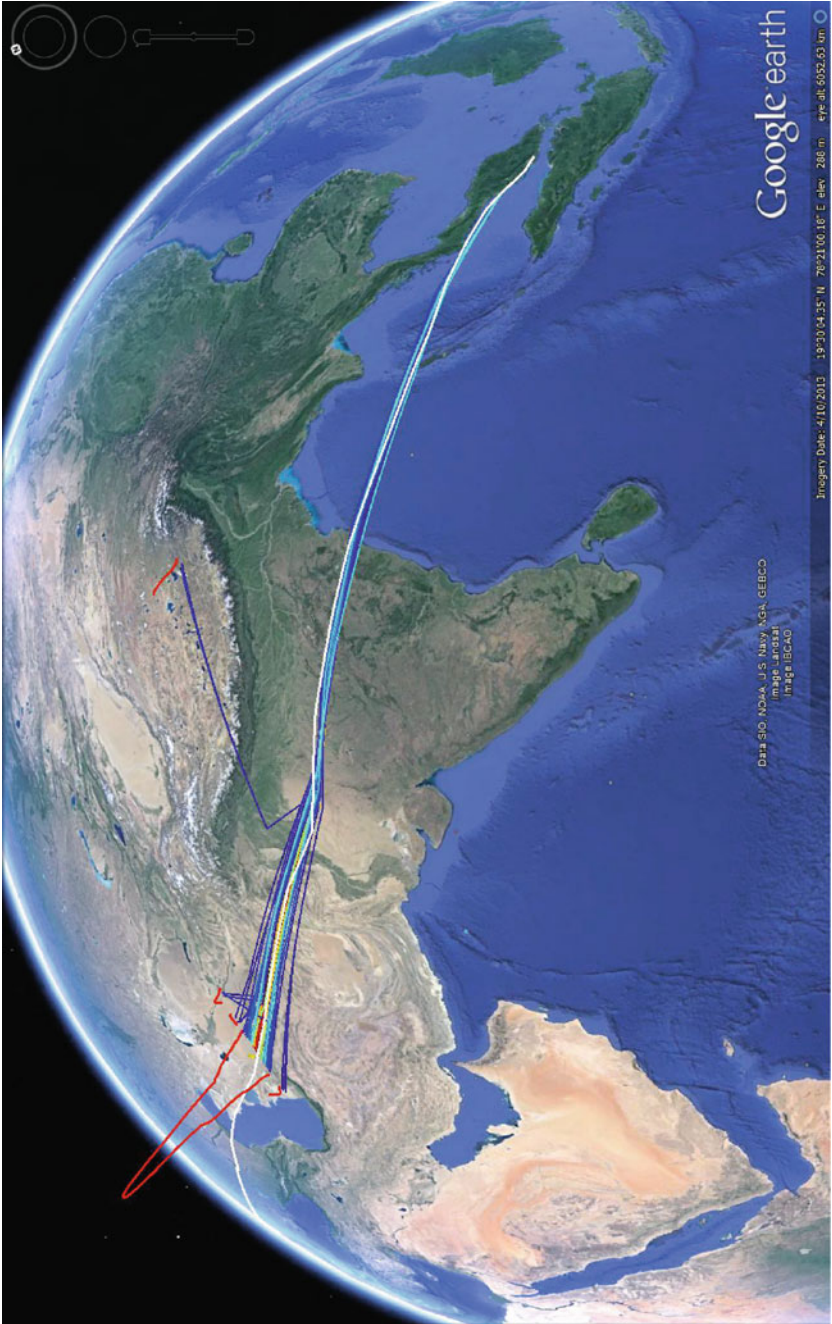
Fig. 9.12 Validation flight 7 March 2014 Kuala Lumpur to Frankfurt

The selection process was repeated 10 times for each flight and these 10 measurement sets are treated as independent Monte Carlo random trials for a fixed true aircraft trajectory. As discussed in Chap. 5, the BFO measurement errors are not truly independent over short time periods, which somewhat compromises the assumed independence. However, the common geometry of multiple sets from a single flight is the dominant source of correlation amongst single-flight predictions.

We now briefly review the method used to select individual messages and the performance measure used for this analysis. The chapter concludes with numerical results from these sixty measurement sets.

9.7.1   Measurement Selection

The start and end time for analysis was manually selected for each flight. These times were chosen to exclude ascent from take-off and descent to landing as well as to avoid turns that were very close to either end point. Once these times were determined, the individual measurements were selected using a heuristic randomised process. The intent of this process was to avoid manual selection bias and to create measurement sets that emulate the data available for the accident flight. Measurements were selected recursively.

Let t_{k-1} denote the measurement time for the previous measurement; t_0 is the manually selected starting time. Each measurement has a collection time labelled t_j, for $j \in \{1, \ldots, J\}$, where J is the total number of measurements in the communication log. The first measurement was selected by assigning a probability

$$p_j(0) = P(0)^{-1} \exp \left\{ -\frac{1}{2\sigma^2} \left(t_j - t_0 \right)^2 \right\}, \tag{9.1}$$

$$P(0) = \sum_{j=1}^{J} \exp \left\{ -\frac{1}{2\sigma^2} \left(t_j - t_0 \right)^2 \right\}, \tag{9.2}$$

where σ was chosen to be 15 min. The selected measurement was then chosen by taking a single multinomial draw on the probability vector $p(0)$. This selection prefers measurements closer to the start time. Subsequent measurements were chosen with a mean time spacing of 1 h. Let $l(i)$ index the measurement chosen as the ith in the sequence. A probability vector for the $(i+1)$th measurement was defined as

$$p_j(i+1) = \begin{cases} P(i+1)^{-1} \exp \left\{ -\frac{1}{2\sigma^2} \left(t_j - t_{l(i)} - 1 \right)^2 \right\}, & j > l(i) \\ 0, & j \leq l(i), \end{cases} \tag{9.3}$$

$$P(i+1) = \sum_{j=l(i)+1}^{J} \exp \left\{ -\frac{1}{2\sigma^2} \left(t_j - t_{l(i)} - 1 \right)^2 \right\}. \tag{9.4}$$
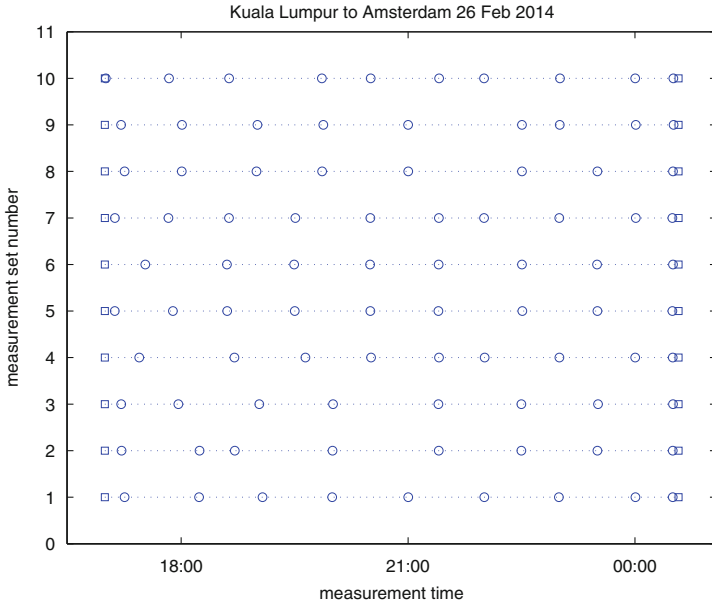
Fig. 9.13 Example measurement timings for a single flight

Measurement $(i + 1)$ is again selected using a single draw on a multinomial distribution defined by $p_j(i + 1)$. The process concludes when the measurement selected occurs after the desired end time: this measurement is discarded.

Figure 9.13 shows an example of the measurement times for the ten different sets generated for the Kuala Lumpur to Amsterdam flight on 26 February 2014. Squares are used to mark the initialisation time and the final time point, neither of which have measurements. The measurement times are marked with circles. Each row is a realisation of the measurement selection process. Some measurements are used by more than one of the sets. The number of measurements selected varies between eight and ten, the duration of the flight segment is approximately seven hours and 35 min: seven one-hour spaces would lead to eight measurements in seven hours.

9.7.2 Performance Measure

In the object tracking literature it is common to use accuracy measures to quantify tracking performance, for example [7]. Accuracy measures quantify how well the estimates from the tracker match the truth. The most frequently used accuracy measure is root-mean-square (RMS) error, which is typically the average geometric distance between the true object position and the tracker estimated position. The requirement for MH370 is a search region, not a point estimate, so RMS is not

applicable. The other common accuracy measure is the Normalised Estimation Error Squared (NEES). This is defined as the expectation of the inner product of the estimation error with itself, normalised by the estimator covariance. For a scalar, this is the mean squared error divided by the filter covariance estimate. Whereas RMS quantifies how accurately the filter finds the centre of mass of a distribution, NEES quantifies how accurately the filter estimates the spread of a distribution. NEES inherently assumes a uni-modal distribution. Again, NEES is based on an assumed Gaussian system with a point estimate and covariance estimate. It is not an appropriate measure for the multi-modal pdf produced by the filter in this application. Instead, the statistical performance of the filter output was quantified by measuring the highest posterior density (HPD) interval at the true aircraft location. The HPD interval is defined as the spatial region for which the filter output pdf is at least as high as the value at the true location. Figure 9.14 shows an example of this process for a scalar random variable x with a Gaussian mixture pdf $p(x)$. The two components are equally weighted, one with mean 2 and variance 0.25 and the other with mean 5 and unit variance. Supposing that the truth in this case was $x = 6$, the HPD interval is shaded in red and corresponds to the regions in x for which $p(x) \geq p(6)$. Because the distribution $p(x)$ has two modes and the value of $p(6)$ is between the lower peak and the intermediate minimum, the HPD is composed of two intervals. If the truth had been 2.5 instead then only one region around the higher peak at 2 would be in the HPD and if the truth were 8 then almost all of the pdf would be in the HPD region.

The integral of the pdf over the HPD interval corresponds to the cumulative probability that a random sample from the distribution is more likely than the truth point. If the integral is close to unity, then the HPD interval contains most of the
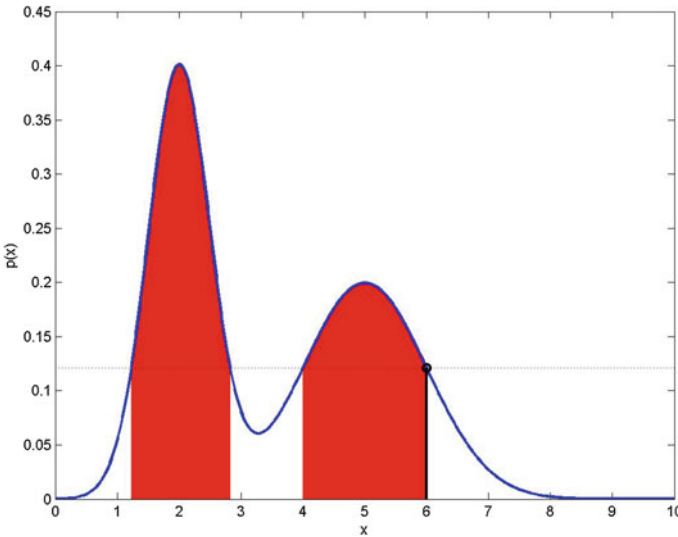
Fig. 9.14 Highest posterior density interval of a Gaussian mixture

support of the pdf, that is the truth point is at a very low part of the pdf. Alternatively, if the integral under the HPD is close to zero then only a small portion of the event space is more likely than the truth point.

Mathematically, the HPD integral is given by

$$h\big(\mathbf{x}^{\text{truth}}; p(\mathbf{x})\big) = \int_{\mathbf{x}:p(\mathbf{x})\geq p(\mathbf{x}^{\text{truth}})} p(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{9.5}$$

where $\mathbf{x}^{\text{truth}}$ is the true aircraft location and $p(\mathbf{x})$ is the filter output pdf. In the discussion that follows, we abbreviate as $h \equiv h\big(\mathbf{x}^{\text{truth}}; p(\mathbf{x})\big)$ the random variable derived by transforming the random variable $\mathbf{x}^{\text{truth}}$ using (9.5).

If the truth values were indeed random samples from the filter output pdfs, then it is relatively easy to show that the distribution of h would be uniform on the interval $[0, 1]$.[1] If integrals tend to be clumped closer to zero then the pdfs being assessed are pessimistic: the tails decay too slowly and the coverage of the pdf is too broad. If the integrals tend to be clumped closer to unity then the truth is always in the tails and the pdfs being assessed are overly optimistic. For the MH370 search definition we prefer a conservative pdf that is a little pessimistic, in order to minimise the chance of excluding the true aircraft location. Provided the search zone defined can be feasibly measured it is better to make this region a little too large and improve the likelihood that the truth is contained.

For each flight we have only ten different measurement sets so it is not feasible to construct a sensible estimate of $p(h)$. Instead we plot an estimate of its cumulative distribution and compare it with the line $y = x$, which is the cumulative distribution of a uniform random variable. If the h values are relatively small then the empirical cumulative distribution function (cdf) will rise more quickly than the reference and the curve will be above it. Conversely if the values are relatively large then the empirical cdf will rise slowly and the curve will be below the reference.

9.7.3   Results

Figure 9.15 shows the empirical cdf derived for each validation flight separately. This shows that the results within a single flight are quite correlated because the filter performance is dependent on geometry. For the Mumbai to Kuala Lumpur, Kuala Lumpur to Beijing and Beijing to Kuala Lumpur flights the h values are generally small but not close to zero. This indicates that the spread of the filter pdf is too high

---

[1]To see this, let $Y = p(\mathbf{x})$, i.e., the random variable obtained by applying the random value $\mathbf{x}$ to its pdf. Then the cumulative distribution function (cdf) of Y, $F_Y(y) = P(Y \leq y)$ is one minus the HPD value in (9.5). It is well-known that, assuming continuity and monotonicity of the cdf, the random variable obtained by passing a random value through its cdf is uniform on the interval $[0, 1]$ (e.g., [13]), and if Y is uniform on $[0, 1]$, then so is $1 - Y$. The necessary assumptions are satisfied if the pdf $p(\mathbf{x})$ contains no non-zero flat regions and no Dirac delta components.
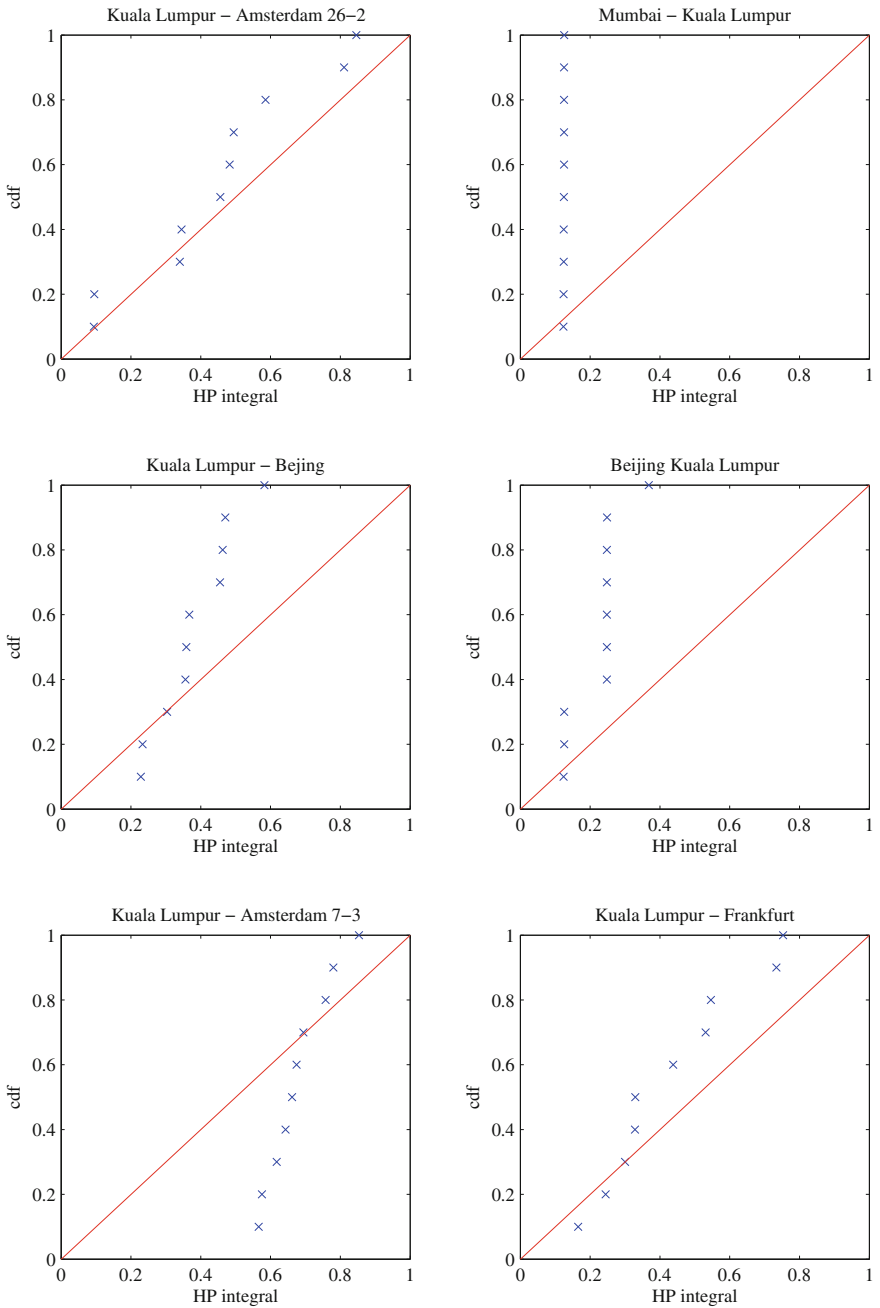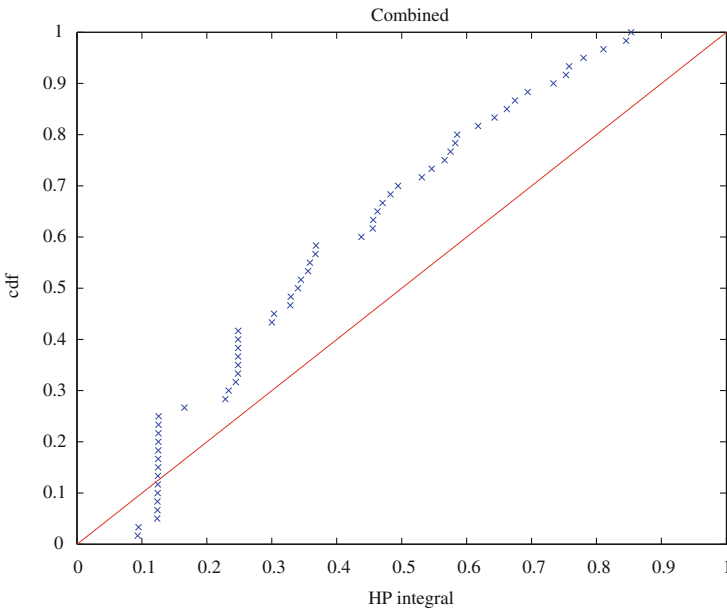
Fig. 9.15 Cumulative density plots for individual validation flights. *Crosses* show individual data-set results and the *solid line* shows the theoretical result for independent samples

and that the peak of the pdf is biased. The bias occurs because the flights make small manoeuvres that are unobservable by the filter. For example, Fig. 9.5 shows that in the Kuala Lumpur to Beijing flight the aircraft made a number of heading changes that lasted for only a short time before the heading reverted back to its previous value. The minor course corrections result in a displacement in position. The filter will sample these paths but their dynamics are less likely than paths without a manoeuvre. For these flights the mode is not a reliable indicator of the true aircraft location but a fairly tight interval is.

In the longer Asia to Europe flights the h values tend to spread between 0.25 and 0.8. Again there is bias due to the repeated geometry and very large values are not observed because the model allows manoeuvres that are more dynamic than what occurred in the actual flights and this spreads the pdf.

Figure 9.16 combines all of the trials into a single h cumulative distribution. In this plot the two different groups of flights are apparent: there is an initial very sharp rise due to the contributions of the intra-Asia flights and then a gradual climb from the Asia to Europe flights.

Overall the results show that for all of the flights and measurement combinations tested the true aircraft location was inside a 85 % confidence region of the pdf. That is, the largest h value observed was approximately 0.85. This means that the pdf estimates are conservative. The spread of the estimated pdf is wider than the spread of true values. This occurs for two reasons: firstly, the aircraft dynamic

Fig. 9.16 Cumulative density plot combined over all validation flights. *Crosses* show individual data-set results and the *solid line* shows the theoretical result for independent samples

model provides more flexibility than is typically used; for example, in most commercial flights, smaller turns are more likely than turns of 90° or more. Secondly, the assumed measurement variances were deliberately inflated to be pessimistic, as discussed in Sect. 5.3. Given that the accident flight was not a typical commercial flight, the dynamic model should not be exactly matched to typical commercial flights. A somewhat conservative pdf in this case is desirable so long as the pdf does not spread over an area that is unreasonably large to search.