Diagnostic plots for dynamical systems

Richard Wilkinson, Kamonrat Suphawan, Theo Kypraios

School of Mathematical Sciences
University of Nottingham

r.d.wilkinson@nottingham.ac.uk

UCM - July 2012

Introduction

- Dynamical systems and common types of error
- Predictive distributions
- Scoring rules
- Diagnostic plots
- Toy example
- Less toyish example

Dynamical systems simulators

$$\begin{tikzcd}
{} \arrow[r] & x_{t-1} \arrow[r] \arrow[d] & x_t \arrow[r] \arrow[d] & x_{t+1} \arrow[r] \arrow[d] & x_{t+2} \arrow[d] \\
& y_{t-1} & y_t & y_{t+1} & y_{t+2}
\end{tikzcd}$$

- State vector x_t which evolves through time. Let $x_{0:T}$ denote $(x_0, x_1, \dots, x_T)$.
- Computer model f which encapsulates our beliefs about the dynamics of the state vector

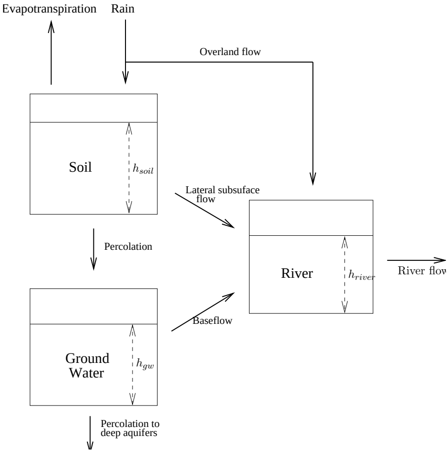
$$x_{t+1} = f(x_t, u_t) + w_t$$

  where w_t represents a simulator discrepancy term (can depend on u and x). Treat f as a black-box
- Observations $y_t = h(x_t)$ where $h(\cdot)$ usually contains some stochastic element

We're not doing parameter estimation.

UCM 2010

Wilkinson *et al.* 2011



Assume

$$x_{t+1} = f(x_t) + \delta(x_t)$$

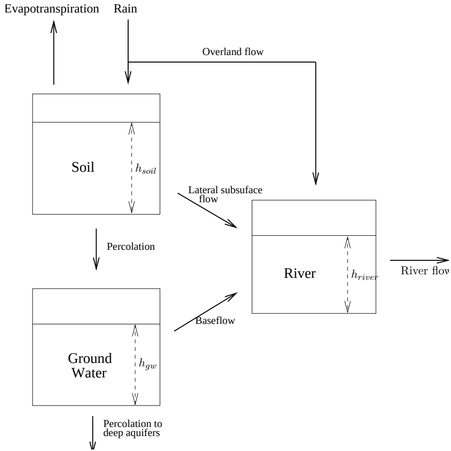
where $\delta(x)$ is a functional simulator discrepancy.

We then proposed a model for $\delta(\cdot)$, parametric or otherwise, and attempted to learn it.

This turned out to be hard - we never observe x_t, only y_t, which may be lower dimensional and only provide limited information about x_t.

UCM 2010

Wilkinson *et al.* 2011



Assume

$$x_{t+1} = f(x_t) + \delta(x_t)$$

where $\delta(x)$ is a functional simulator discrepancy.

We then proposed a model for $\delta(\cdot)$, parametric or otherwise, and attempted to learn it.

This turned out to be hard - we never observe x_t, only y_t, which may be lower dimensional and only provide limited information about x_t.

Simpler question: Given an imperfect statistical forecasting system, diagnose which aspect of the system is causing the error

All models are wrong...

The forecasting system consists of a simulator, statistical model of simulator discrepancy, observation equation and statistical model, inferential scheme... call the entire system a *statistical forecasting system*

There are various types of error we might see

- Incorrect simulator dynamics (simulator discrepancy)
- Simulator discrepancy term mis-specified
- Measurement process mis-specified (incorrect variance, incorrect h)
- Initial condition errors - poor choice of x_o when initialising forecasts

  $\vdots$

Given an imperfect forecasting system, how do we know what type of error we are faced with?

- we've looked in various literatures, but found little

System output

We must decide what aspect of the system output to use, and then how to judge it.

Possible system outputs:

- Posterior predictive distributions $\pi(y_t^{rep}|y_{1:t})$ or smoothed distributions $\pi(y_t|y_{1:T})$ for $1 \leq t \leq T$
- Predictive distributions $\pi(y_{t+k}|y_{1:t})$ - k-step-ahead forecast
- Point estimates from any of the above

Note: obtaining these distributions typically requires some work, e.g. Kalman filter and its variants, sequential Monte Carlo methods, ABC methods. (needed?)

System output

We must decide what aspect of the system output to use, and then how to judge it.

Possible system outputs:

- Posterior predictive distributions $\pi(y_t^{rep}|y_{1:t})$ or smoothed distributions $\pi(y_t|y_{1:T})$ for $1 \leq t \leq T$
- Predictive distributions $\pi(y_{t+k}|y_{1:t})$ - k-step-ahead forecast
- Point estimates from any of the above

Note: obtaining these distributions typically requires some work, e.g. Kalman filter and its variants, sequential Monte Carlo methods, ABC methods. (needed?)

posterior predictive distributions: $\exists$ a debate over their validity

- Difficulty of interpretation - conditioning on the data moves the state prediction closer to the data, making probabilities harder to interpret.
- p-values don't have a $U[0,1]$ distribution under the null
- Prior predictive p-values have been recommended instead by various authors (e.g., Bayarri and Castellanos 2007)

k-step ahead-predictive distributions

In the dynamical systems setting, the prior predictive density is the k-step-ahead forecast $\pi(y_{t+k}|y_{1:t})$

Prediction vs explanation (Shmueli 2011)

- base validation and diagnostics solely on the forecasting system's abilty to predict, not explain
- Link to Dawid's prequential approach
- over-fitting less of a problem if we are only using predictive measures

k-step ahead-predictive distributions

In the dynamical systems setting, the prior predictive density is the k-step-ahead forecast $\pi(y_{t+k}|y_{1:t})$

Prediction vs explanation (Shmueli 2011)

- base validation and diagnostics solely on the forecasting system's abilty to predict, not explain
- Link to Dawid's prequential approach
- over-fitting less of a problem if we are only using predictive measures

By looking at predictions at different lead times we can emphasise different aspects of the forecasting system:

$$\begin{tikzcd}
{} \arrow[r] & x_t \arrow[r, "+\sigma^2"'] \arrow[d] & x_{t+1} \arrow[r, "+\sigma^2"'] \arrow[d, "+\tau^2"] & x_{t+2} \arrow[r, "+\sigma^2"'] \arrow[d] & x_{t+3} \arrow[d, "+\tau^2"] \\
 & y_t & y_{t+1} & y_{t+2} & y_{t+3}
\end{tikzcd}$$

How to judge: Numerical Scores

Numerous numerical scores are used e.g., MSE, MAD, MAPE, and corresponding skill-scores obtained by comparing these values to those given for a reference forecasting system such as climatology or persistence, e.g., Nash-Sutcliffe efficiency, Theil's U.

How to judge: Numerical Scores

Numerous numerical scores are used e.g., MSE, MAD, MAPE, and corresponding skill-scores obtained by comparing these values to those given for a reference forecasting system such as climatology or persistence, e.g., Nash-Sutcliffe efficiency, Theil's U.

Theory of proper scoring rules undergoing a resurgence in recent years. A scoring rule S takes forecast distribution π and observed value y and returns a numerical score $S(\pi, x) \in \mathbb{R}$

S is proper ifF

$$\mathbb{E}_q S(\pi, X) = \int S(\pi, x) q(\mathrm{d}x)$$

is maximised at $\pi = q$. That is, if we believe q, then we maximize our expected score by reporting q.

How to judge: Numerical Scores

Numerous numerical scores are used e.g., MSE, MAD, MAPE, and corresponding skill-scores obtained by comparing these values to those given for a reference forecasting system such as climatology or persistence, e.g., Nash-Sutcliffe efficiency, Theil's U.

Theory of proper scoring rules undergoing a resurgence in recent years. A scoring rule S takes forecast distribution π and observed value y and returns a numerical score $S(\pi, x) \in \mathbb{R}$

S is proper ifF

$$\mathbb{E}_q S(\pi, X) = \int S(\pi, x) q(\mathrm{d}x)$$

is maximised at $\pi = q$. That is, if we believe q, then we maximize our expected score by reporting q.

Important to use proper scores, as improper scores are an inducement to hedging

Examples include the CRPS, logarithmic score, Brier score, Dawid score...

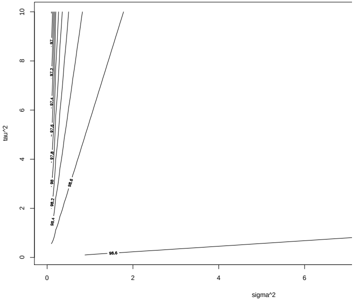
Any proper score can be decomposed into a reliability term and a sharpness term.

Improper score vs proper score

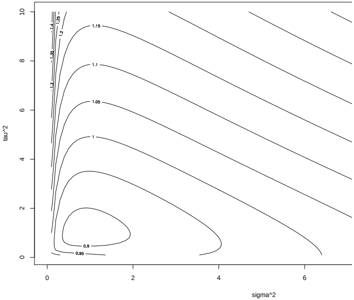
Two parameter model, both scale parameters, 1-step ahead forecasts

- Simulator discrepancy variance on x-axis - true value at 1
- Measurement error variance on y-axis - true value at 1
- Score on z-axis

MSE contour surface



CRPS contour surface



CRPS correctly identifies region containing true value
MSE can only find good ratios of parameter values.

Diagnostic plots

We want to be able to say where the problem lies, not just that there is a problem.

Testing and formal inference are hard problems for state-space problems because of the high-dimensional nature of the missing data.

Tukey (1962, and many others) suggested we use graphical methods to highlight problems, rather than parametrizing types of departure in advance and developing significance tests.

Diagnostic plots seem to be a better option than deterministic scores. Useful plots include

- Residuals $\epsilon_{t+k} = y_{t+k} - \mathbb{E}(y_{t+k}|y_t)$ vs time $t+k$
- Residuals ϵ_{t+k} vs posterior mean $\mathbb{E}(y_{t+k}|y_t)$
- Autocorrelation (ACF) plots at various lags
- CUSUM type plots.

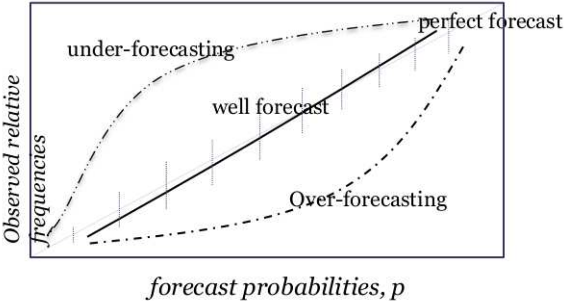
Attribute diagrams

Hsu and Murphy 1986

- Sequence of binary events with outcomes $d_1, d_2, \ldots, d_T$
- Sequence of forecast probabilities $f_1, f_t, \ldots, f_T$.

Partition $[0, 1]$ into K intervals with each interval characterized by a representative forecast probability $\bar{f}_k$.

- let $\bar{d}_k$ be the relative occurence of the event when the forecast was in interval k.
- Attribute diagram is the plot of $\bar{d}_k$ vs $\bar{f}_k$.
- We can add simple error bars using the Monte Carlo variance.

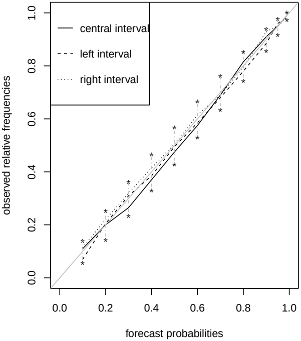


Attribute diagrams for continuous random variables

Attribute diagrams are defined for sequences of binary events.
We artificially create such a sequence by defining prediction intervals I_t

- central interval $I_t^{(c)} = [a_t, b_t]$
- left interval $I_t^{(l)} = (-\infty, l_t]$
- right interval $I_t^{(r)} = [r_t, \infty)$

so that $\mathbb{P}(y_{t+k} \in I_{t+k}^{(\cdot)} | y_{1:t}) = p$ and then define the sequence of binary events to be



$$r_t^{(p)} = \begin{cases} 1 \text{ if } y_{t+k} \in I_{t+k}^{(\cdot)} \\ 0 \text{ otherwise} \end{cases}$$

We can then compare the relative frequency $\overline{r}^{(p)}$ with p (perfectly calibrated forecast has $\overline{r}^{(p)} = p$).

Repeat for various p and plot to get a version of the attribute diagram for continuous random variables.

Illustrative example 1

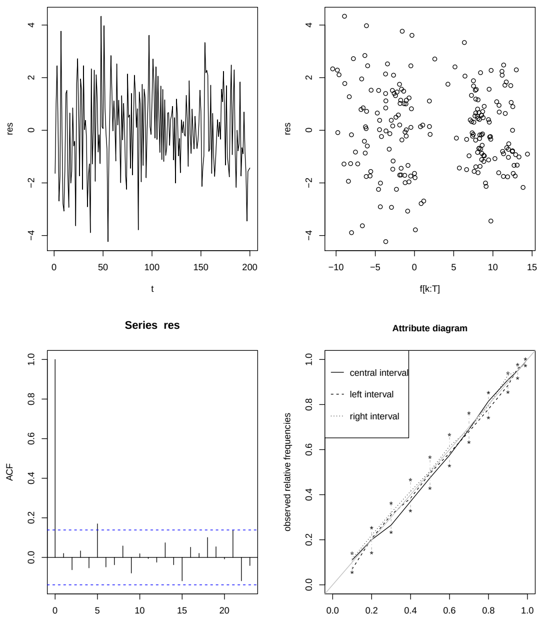
To illustrate the patterns we expect to see for the various diagnostic plots as errors of different types of error are introduced, consider the simple linear Gaussian model

$$X_{t+1} = aX_t + b + N(0, \sigma^2)$$
$$Y_t = cY_t + N(0, \tau^2)$$

We can use the Kalman filter to obtain the filtering distributions and the k-step ahead predictions in this case.

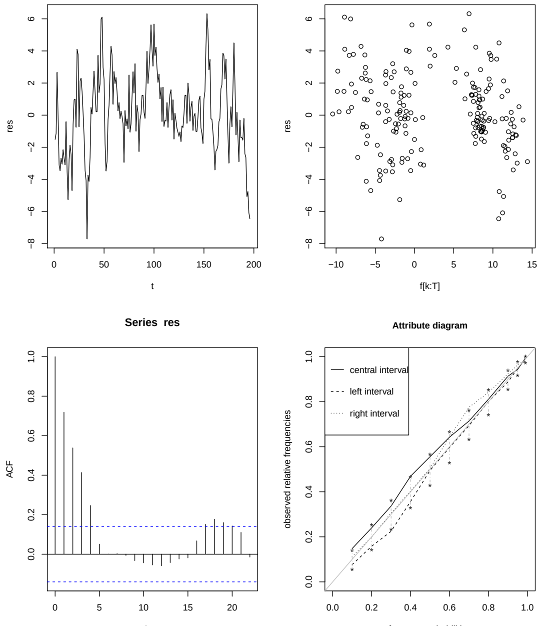
We generate an “observed” dataset with $a = b = c = \tau = \sigma = 1$ and then introduce errors one at a time to illustrate the deviations we expect from the null form of the various plots.

Null plots - 1 step ahead



- Trace plots should look like white noise
- Residual plots a band of points with no noticeable pattern
- ACF plot show no correlaton at lag greater than 1
- Attribute diagrams show no significant departiure from $y = x$

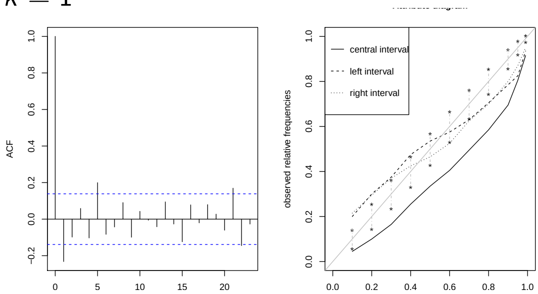
Null plots - 5 step ahead



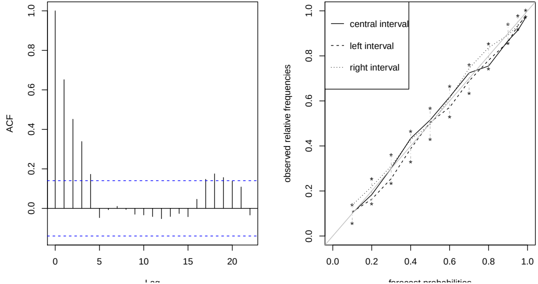
- We now expect to see some correlation as the residuals from predictions less than 5 time points apart will correlated
- ACF plot shows positive correlation for lags upto 5, and then a period of negative correlation.
- Attribute diagrams show no significant departiure from $y = x$

Incorrect measurement error - too small

$k = 1$



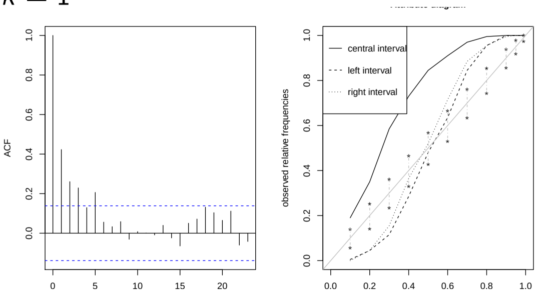
$k = 5$



- Lag 1 correlation will be negative
- Atribute diagram: central interval - over-confident for central interval. For L/R intervals, under-confident for $p < 0.5$ and over-confident for $p > 0.5$.
- As the lead time k increases, the ratio $\frac{k\sigma^2+\tau^2}{k\sigma^2+\tau^2_{true}} \rightarrow 1$, and so the attribute diagrams will begin to look OK

Incorrect measurement error - too large

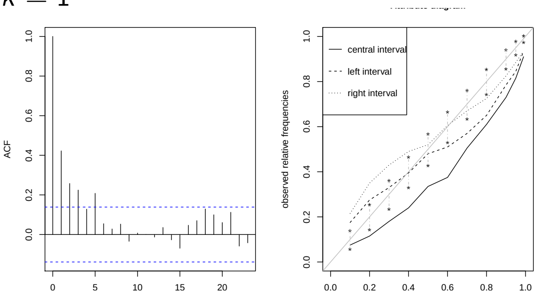
$k = 1$



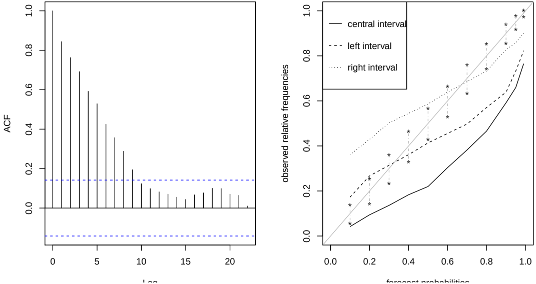
- With τ^2 too large, the pattern is reversed.
- Central interval will be under-confident.
- L/R intervals will be over-confident for $p < 0.5$, otherwise under-confident.
- Attribute diagram looks better as lead time increases.

Incorrect simulator discrepancy error - too small

$k = 1$



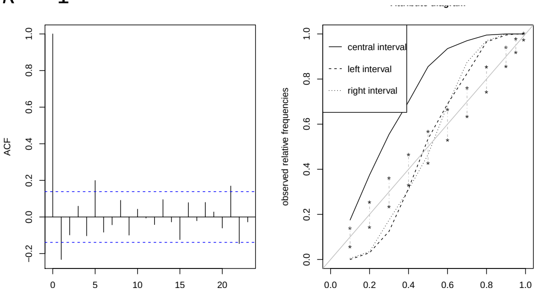
$k = 10$



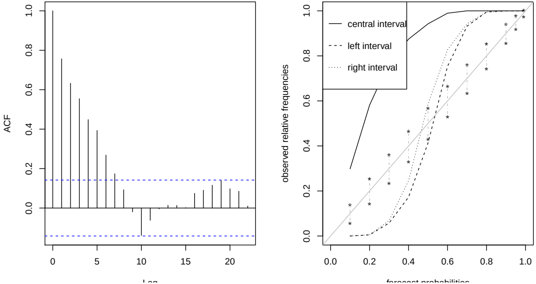
- ACF plot will tend to be positive - too much emphasis on model
- Atribute diagram shows same pattern as when meas. error is too small.
- However, as the lead time k increases, the ratio $\frac{k\sigma^2+\tau^2}{k\sigma^2+\tau^2_{true}} \longrightarrow 1$ slowly increases, and so the attribute diagrams don't improve (and may look worse).

Incorrect simulator discrepancy error - too large

$k = 1$



$k = 10$



- Negative correlation at lag 1 - residuals osciallate (chasing the data)
- Atribute diagram shows same pattern as when meas. error is too large, but the error doesn't improve with lead time.

Incorrect simulator dynamics

Harder to deal with, as the variety of possible errors is larger.

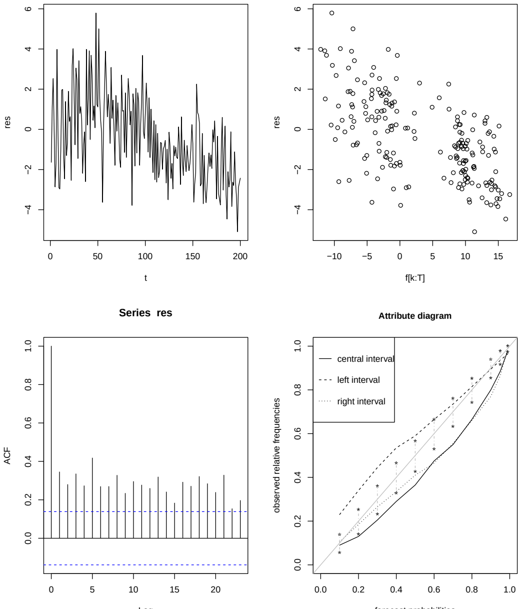
Similar to having too small a simulator discrepancy term.

- Central interval attribute diagrams will be over-confident.

However, there are some differences we may be able to spot.

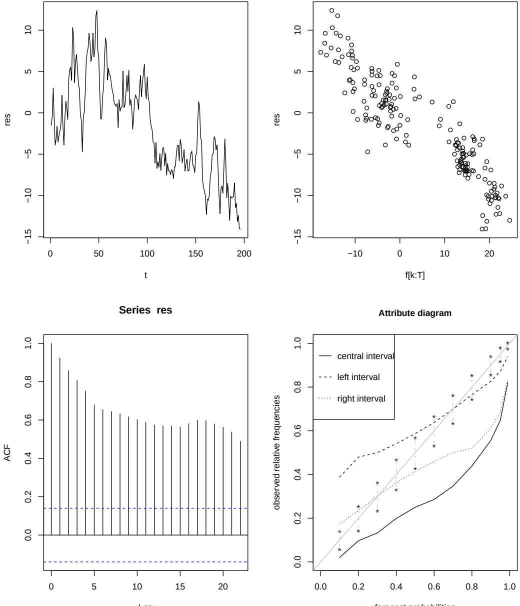
- Patterns of correlation in the residual plots
- ACF plot positive at all lags
- Error grows faster with longer lead times
- A difference evident between the left and right attribute diagram curves.

Incorrect simulator dynamics: $a = 1.1$, lead time $k = 1$



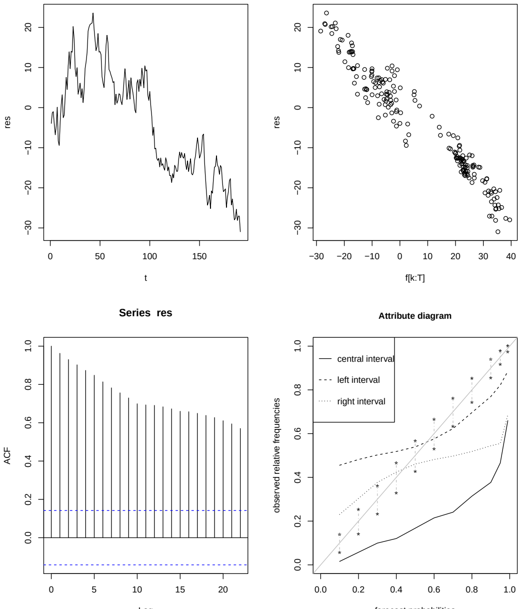
- Residual plot starts to show the shape of the missing dynamics
- L/R attribute diagrams show a difference
- All ACFs positive

Incorrect simulator dynamics: $a = 1.1$, lead time $k = 5$



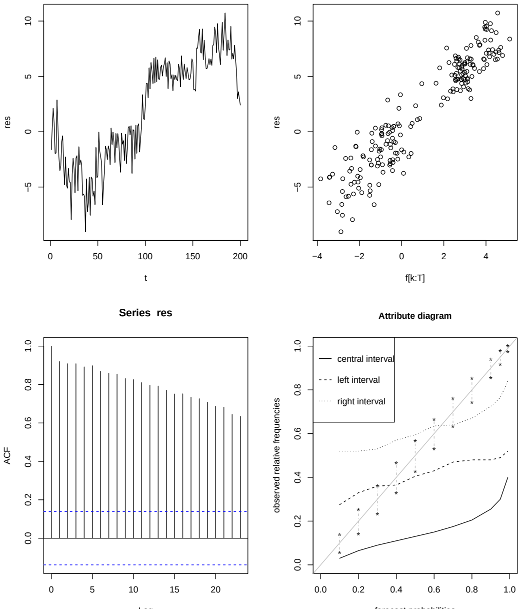
- Residual plot pattern becomes more obvious with longer lead time
- L/R/C attribute diagrams get worse
- L/R interval attribute curves are different
- ACFs large and positive.

Incorrect simulator dynamics: $a = 1.1$, lead time $k = 10$



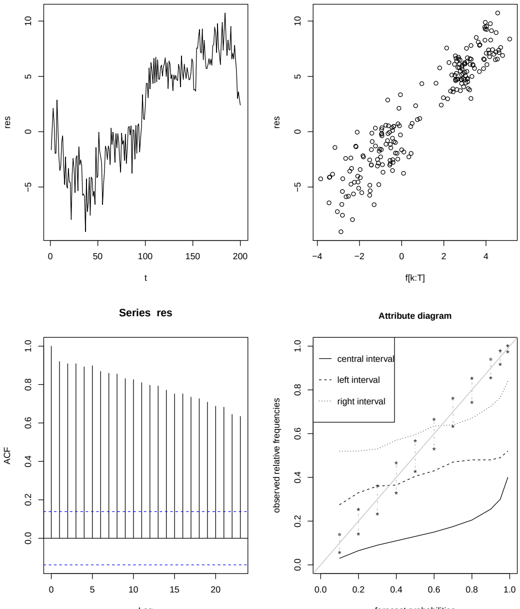
- Large difference between L/R attribute diagrams

Incorrect simulator dynamics: $a = 0.5$, lead time $k = 1$



- Larger simulator error easier to spot

Incorrect simulator dynamics: $a = 0.5$, lead time $k = 1$

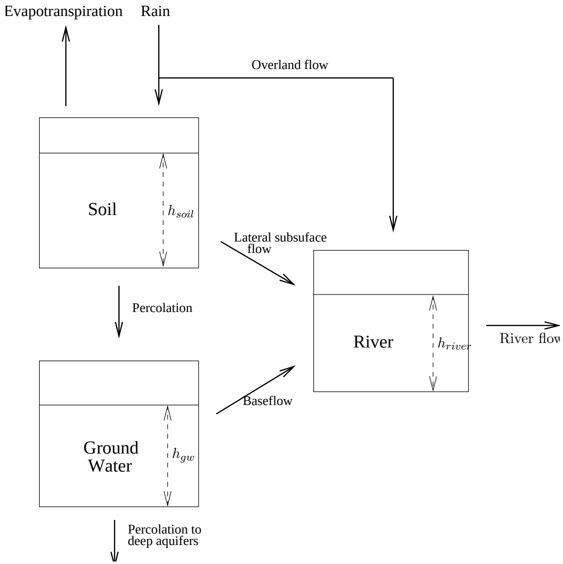


- Larger simulator error easier to spot

If there are multiple types of error, plots give mixed messages. Patterns seem to exhibit behaviour corresponding to dominant error.

Rainfall-runoff simulator

Data from the Abercrombie Valley, Aus



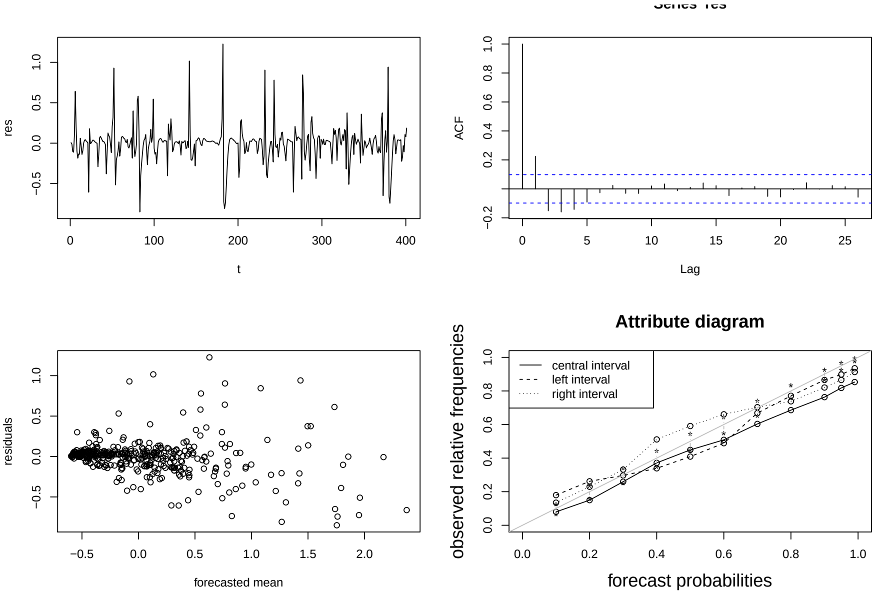
We have measurements of the river flow, evopotranspiration, and rainfall for a 2 year period.

The measurement error process is unknown, but we've assumed

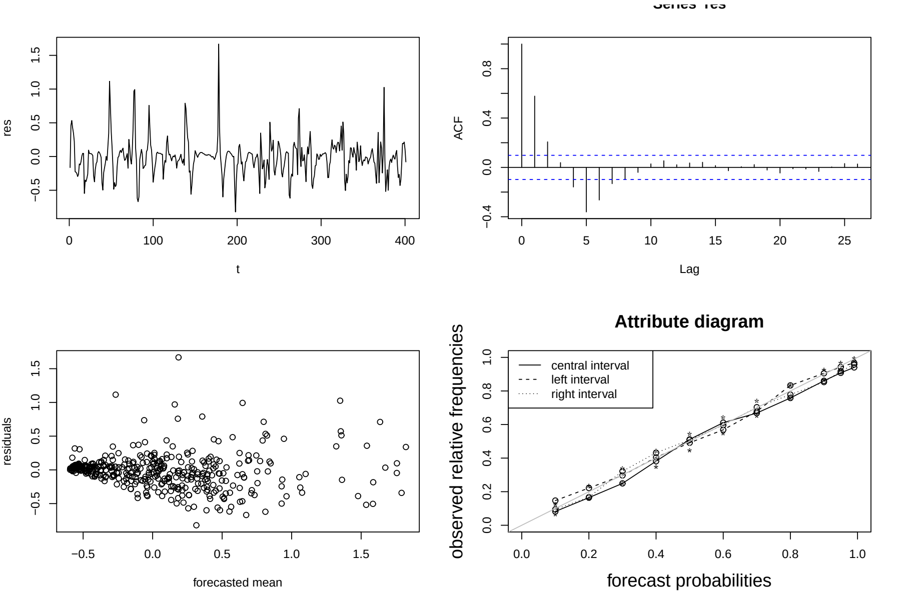
$$\log(R(t)+\lambda) \sim N(\log(R(t)^{sim}+\lambda), s^2)$$

where $R(t)$ is the river flow measurement, and $R(t)^{sim}$ is the simulator prediction + discrepancy. Discrepancy estimated by ML previously.

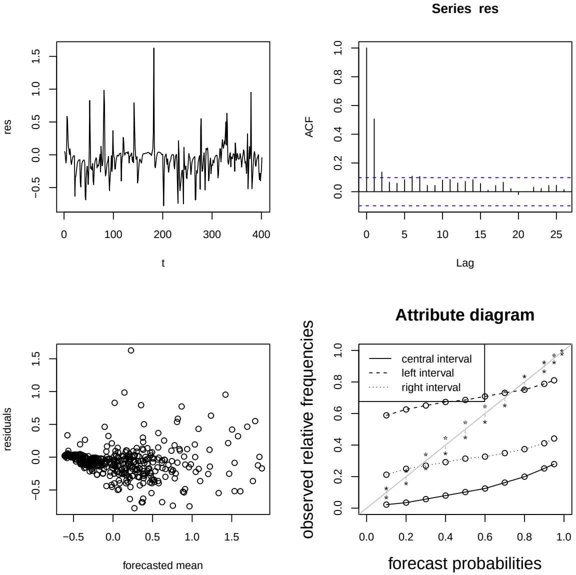
$k = 1$, estimated discrepancy



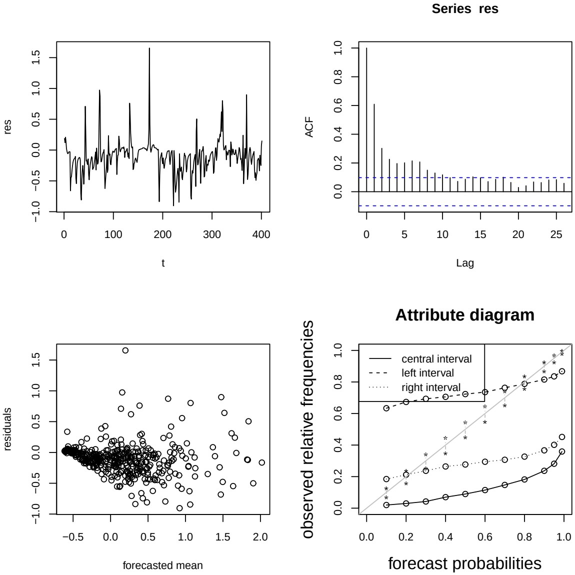
$k = 5$, estimated discrepancy



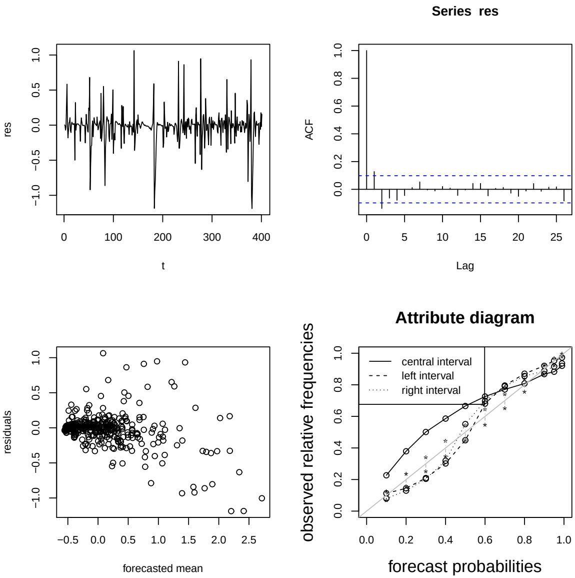
$k = 10$, estimated discrepancy/3



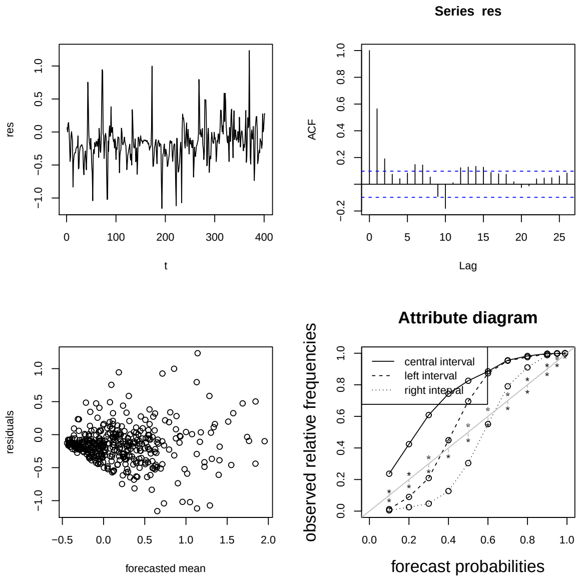
$k = 1$, estimated discrepancy/3



$k = 1$, estimated discrepancy×1.7, measurement error$/3$



$k = 10$, estimated discrepancy$\times 1.7$, measurement error$/3$



Conclusions

Parametrizing all errors and estimating parameter values is fraught with difficulties.

We think a case can be made for using simpler diagnostic tools, at least in the early stage of modelling.

- Predictive distributions more useful for diagnostics than posterior-predictive distributions (forecasts not hindcasts)
- Proper scoring rules needed if scale parameters are unknown
- Using different lead times emphasises different aspects of the problem
- Diagnostic plots more useful than numerical summaries for diagnosing errors
    - Attribute diagrams with artificial prediction intervals can be useful.

However, it seems to be inherently difficult to spot the source of errors, particularly if they are small unless long time-series are available (computational resource?).

Thank you for listening!

To do

- Apply to more real examples
- Combinations of different types of errors - how do we spot and disentangle
- Theoretical properties
  - Prove expected sign of lags
  - Prove expected other patterns in attribute diagrams etc
  - Problems with correlation/dependence in the attribute plots - are the confidence bands correct?
  - Something doesn't seem quite right. As we let T increase, the L/R interval attribute diagrams don't seem to work
  - Theoretically, L/R curves should be the same when no simulator bias, only incorrect simulator or measurement error.
  - Even with true parameter values, the residuals plots doesn't look like a null plot for large lags - eg lag 100 shows clear trend. Why? Is it correlation in the residuals - should we thin to 1/k?