Practical 1: modelling climate change

- What is the rate of temperature increase at Cambridge?
  (See section 2.2.5 from lecture notes.)

- Are temperatures increasing at a constant rate, or has the increase accelerated?
  (See questions 8 and 9 on example sheet 1.)

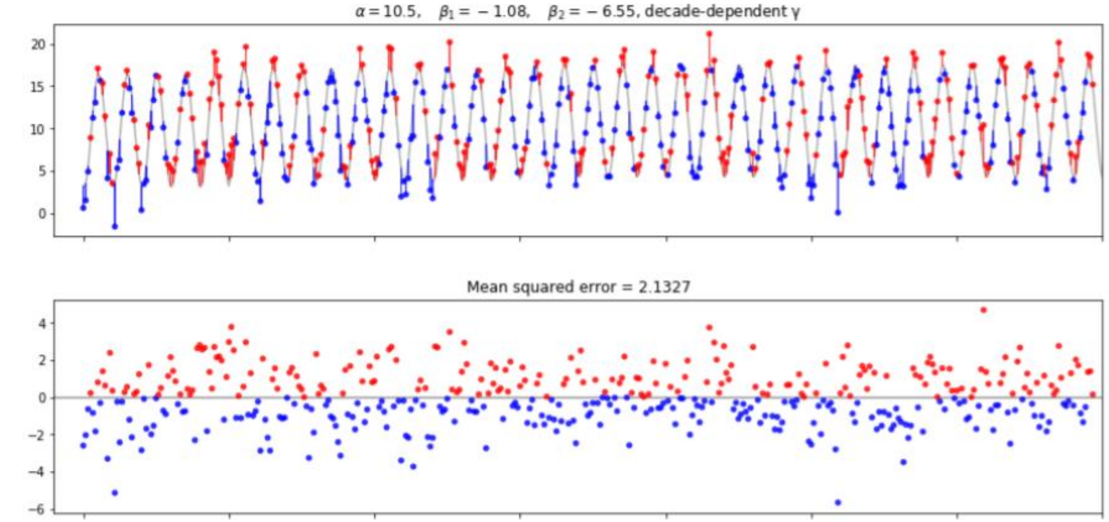
- How do the results compare across the whole of the UK?
  (Use one-hot coding for this, don't iterate over individual weather stations.)

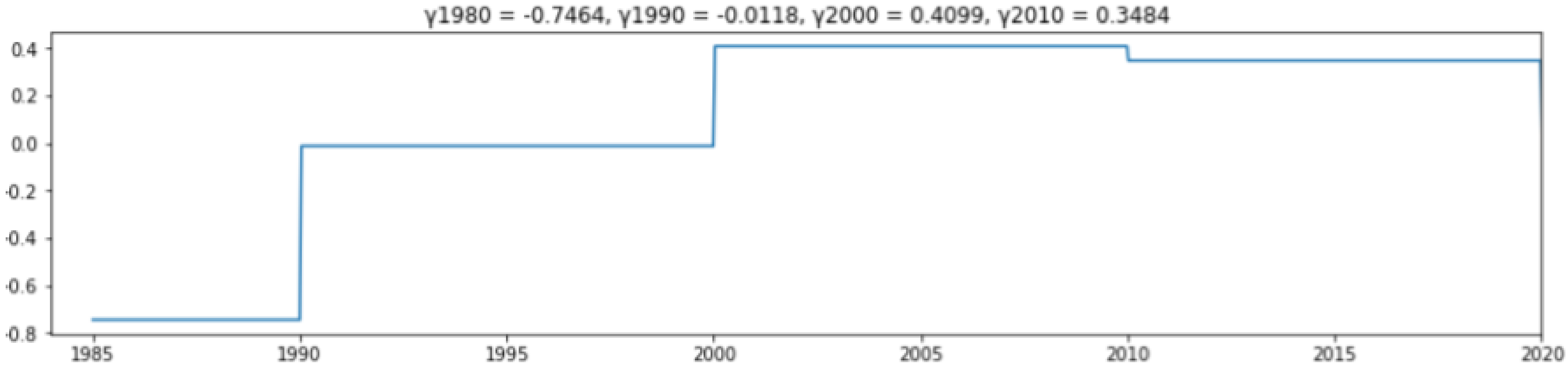
Your task is to produce plots that communicate findings from appropriate linear models.

Michelle Wan

Data: Cambridge.
Model:

$$\texttt{temp} \approx \alpha + \beta_1 \sin(2\pi t) + \beta_2 cost(2\pi t) + \gamma_{1980} 1_{t\in 1980s} + \gamma_{1990} 1_{t\in 1990s} + \gamma_{2000} 1_{t\in 2000s} + \gamma_{2010} 1_{t\in 20100s}$$

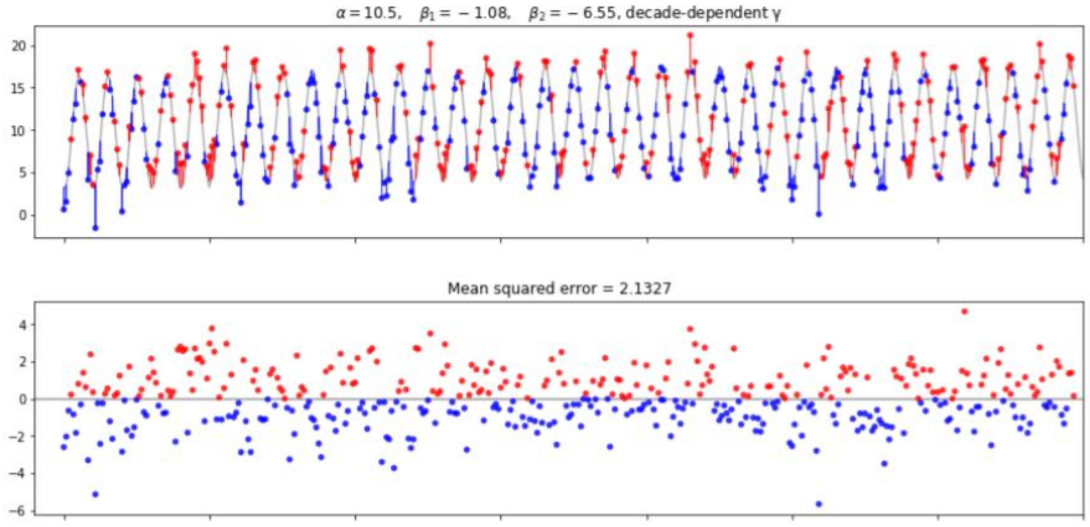




Michelle Wan

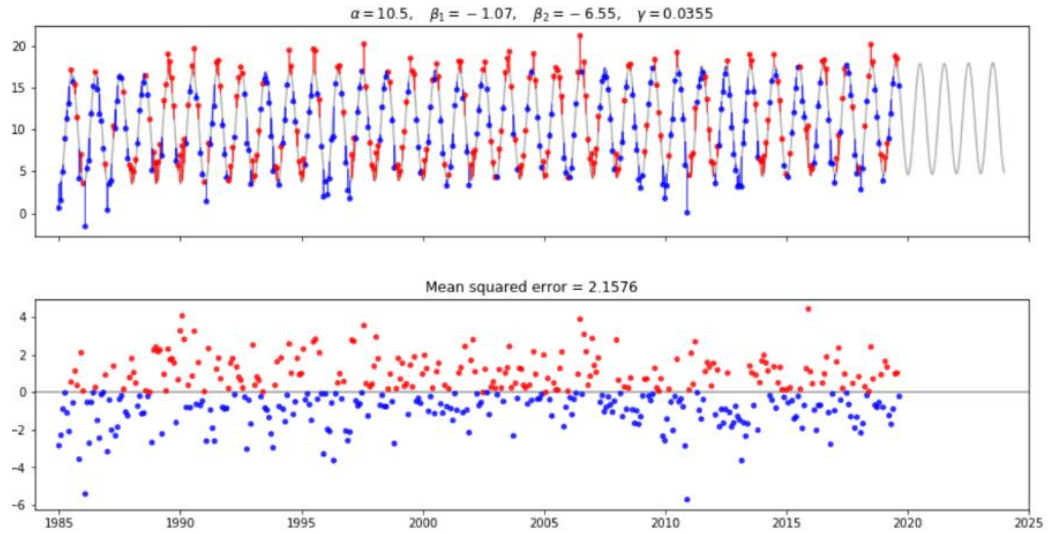
Data: Cambridge.
Model:

$$\texttt{temp} \approx \alpha + \beta_1 \sin(2\pi \texttt{t}) + \beta_2 cost(2\pi \texttt{t}) + \gamma_{1980} 1_{\texttt{t}\in 1980s} + \gamma_{1990} 1_{\texttt{t}\in 1990s} + \gamma_{2000} 1_{\texttt{t}\in 2000s} + \gamma_{2010} 1_{\texttt{t}\in 20100s}$$



Data: Cambridge.
Model:

$$\texttt{temp} \approx \alpha + \beta_1 \sin(2\pi \texttt{t}) + \beta_2 cost(2\pi \texttt{t}) + \gamma(t - 2000)$$



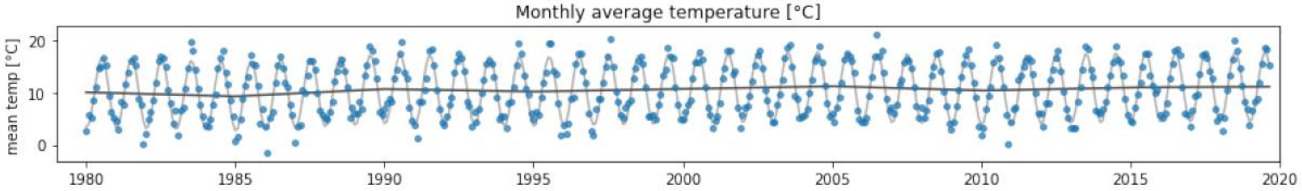
Difference in MSE for the two secular models (decade-independent and decade-dependent) is very small relative to the scale of the temperature data – may suffice to the use the decade-independent model as this has far fewer parameters.

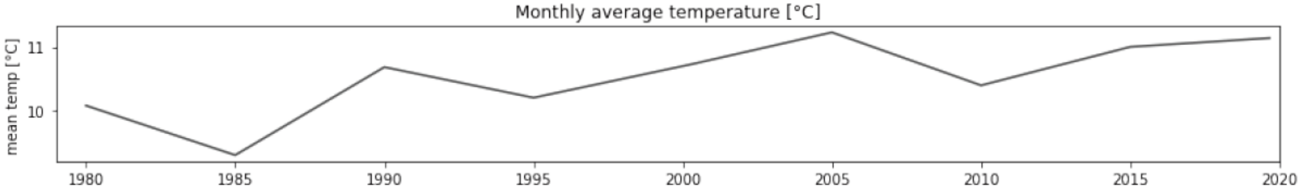
Justin Hou

Data: Cambridge.
Model:

$$\mathtt{temp} \approx \alpha + \beta_1 \sin(2\pi \mathtt{t}) + \beta_2 cost(2\pi \mathtt{t}) + \sum_i m_i 1_{\mathtt{t} \geq \mathrm{base}_i}(\mathtt{t} - \mathrm{base}_i)$$

- The 'buckets' are reduced from decades to 5 years to hopefully give a better 'resolution'.
- Starting the year range for the dataframe from 1985, gives an abnormally high value for m_{i}, which seems to be resolved if the data starts from 1980.



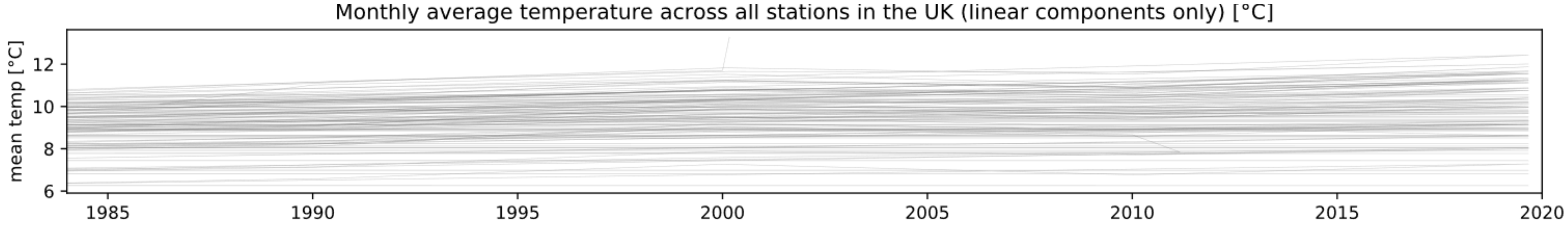


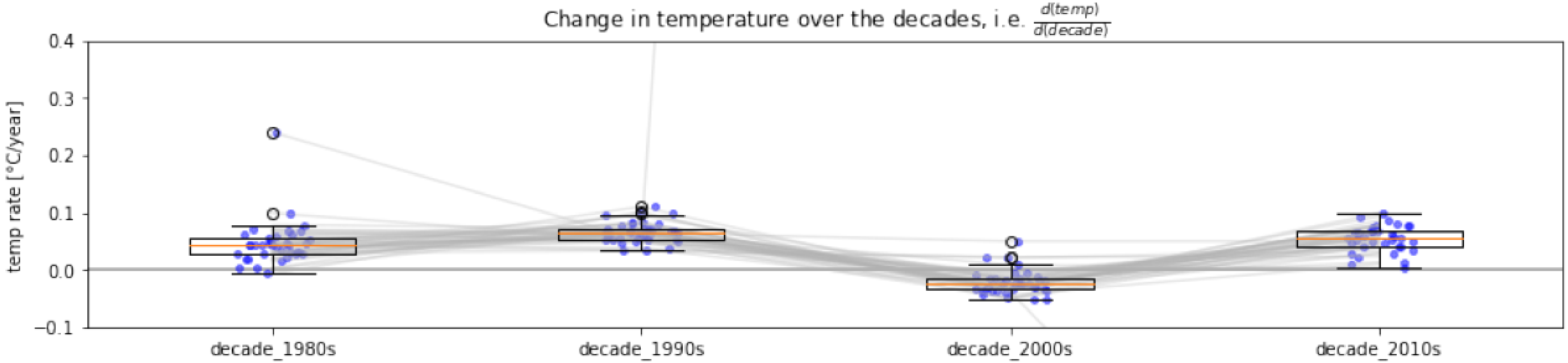
$\mathrm{base}_0 = 1980$ $\mathrm{base}_1 = 1985$ $\mathrm{base}_2 = 1990$

Justin Hou

Data: UK.
Model:

$$\texttt{temp} \approx \alpha_{\texttt{stat}} + \beta_{1,\texttt{stat}} \sin(2\pi \texttt{t}) + \beta_{2,\texttt{stat}} cost(2\pi \texttt{t}) + \sum_i m_{i,\texttt{stat}} 1_{\texttt{t} \geq \text{base}_i} (\texttt{t} - \text{base}_i)$$

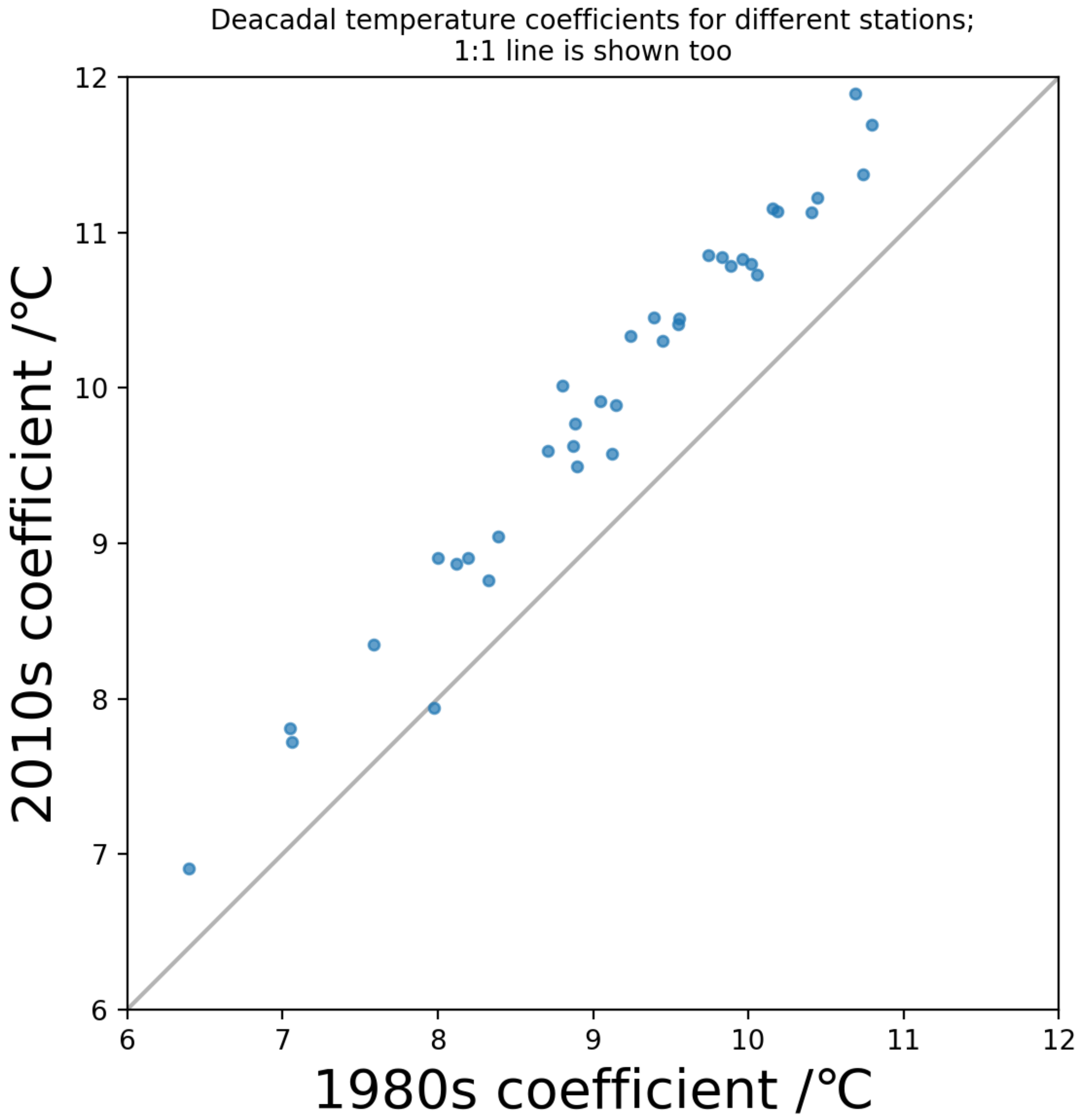




I was not able to draw any clear conclusion from the dataset.

Raghul Parthipan

Data: from all stations
Model: includes a γ_d term, one for each decade d.



Take-aways

- ## Model comparison (Michelle Wan)
  Often we do data science to decide which model to use to describe our data. Does the data push us towards a richer more complicated model, or is a parsimonious model (i.e. one with few parameters) sufficient?

- ## Rich output (Justin Hou)
  When you produce comprehensive outputs, allowing the viewer to see everything in context and to draw their own conclusions, your work will come across as authoritative.

- ## Go meta (Raghul Parthipan)
  Models let you pick out parameters of interest, e.g. “rate of temperature increase”. You can then plot just these parameters. You’ll be able to get your message across more clearly, without it being cluttered with lots of other stories.

- ## The difficulty of analysis without a hypothesis
  Start with a hypothesis you’d like to test, and figure out how to get the data to confirm or refute it. If you start by “I just want to explore this dataset”, you’ll get mired in a swamp of umpteen different models.