Rglimclim: a multivariate, multisite daily weather generator for climate change impact studies

Richard Chandler

r.chandler@ucl.ac.uk

Department of Statistical Science, University College London

SWG 2014, Avignon, September 2014

Research funded by NERC Changing Water Cycle Programme

NATURAL ENVIRONMENT RESEARCH COUNCIL

UCL

Motivating example

- HydEF project (http://www.bgs.ac.uk/changingwatercycle/hydef.html) looking at hydro(geo)logical impacts of climate change in UK
- Detailed hydro(geo)logical models require high-resolution weather inputs, consistent with changing large-scale synoptic conditions as obtained e.g. from reanalysis products or GCMs

Motivating example

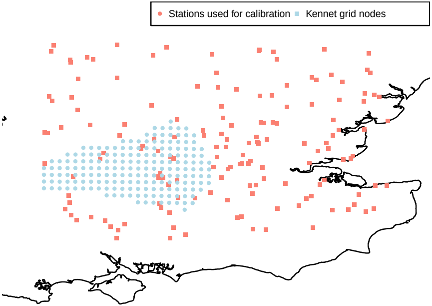
- HydEF project (http://www.bgs.ac.uk/changingwatercycle/hydef.html) looking at hydro(geo)logical impacts of climate change in UK
- Detailed hydro(geo)logical models require high-resolution weather inputs, consistent with changing large-scale synoptic conditions as obtained e.g. from reanalysis products or GCMs

E.g. variables needed by JULES:

Rainfall rate	*Air pressure*	*Snowfall rate*	*Air temperature*
Wind speed	*Specific humidity*	*Downward short-wave radiation*	*Downward long-wave radiation*

Case study: the Thames

- Largest catchment in UK ($\sim$ 10000km^2)
- Modellers wanted hourly sequences, 8 variables, 1km^2 resolution throughout catchment



- Negotiated settlement: daily sequences, 5×5km^2 resolution, Kennet subcatchment (186 grid nodes)
- Data on (most) variables nominally available from 157 stations, 1970 onwards

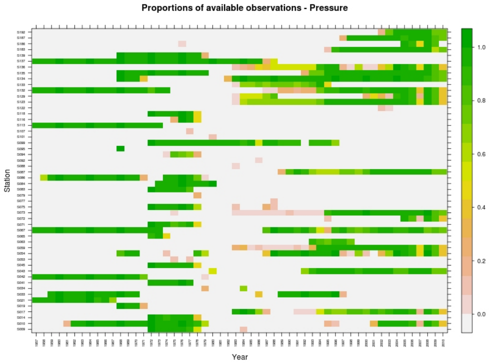
Data availability (I)

- Hourly data obtained from British Atmospheric Data Centre (BADC), MIDAS Met Office dataset
- Available variables: rainfall, snow, air pressure, air temperature, wind speed, downward SW radiation
- Missing variables: specific humidity and downward LW radiation
    - Can be derived from other variables using standard procedures from literature
- **BUT** . . .

Data availability (II)

Numbers of stations with data (out of 157)

Rainfall	Pressure	Temperature	Wind speed	SWR
71	52	140	135	22



- Many stations have short / incomplete/ patchy records

Weather generator requirements

- Need to generate daily data for . . .

Weather generator requirements

- Need to generate daily data for . . .
- Several variables simultaneously, with different distributions and preserving inter-variable relationships . . .

Weather generator requirements

- Need to generate daily data for . . .
- Several variables simultaneously, with different distributions and preserving inter-variable relationships . . .
- at many locations simultaneously, preserving inter-site relationships . . .

Weather generator requirements

- Need to generate daily data for ...
- Several variables simultaneously, with different distributions and preserving inter-variable relationships ...
- at many locations simultaneously, preserving inter-site relationships ...
- ... including locations for which no observations are available ...

Weather generator requirements

- Need to generate daily data for . . .
- Several variables simultaneously, with different distributions and preserving inter-variable relationships . . .
- at many locations simultaneously, preserving inter-site relationships . . .
- . . . including locations for which no observations are available . . .
- . . . and substantial amounts of missing data at locations where observations *are* available

An approach based on generalised linear models

- **Idea**: model each variable in turn, in each case conditioning on previously-considered variables
  - **Justification**: generalised multiplication law — any joint distribution $f(y_1, y_2, \ldots, y_k)$ can be factorised as

$$f(y_1, y_2, \ldots, y_k) = f_1(y_1) f_2(y_2 | y_1) \ldots f_k(y_k | y_1, \ldots, y_{k-1}) \, .$$

An approach based on generalised linear models

- **Idea**: model each variable in turn, in each case conditioning on previously-considered variables
  - **Justification**: generalised multiplication law — any joint distribution $f(y_1, y_2, \ldots, y_k)$ can be factorised as

$$f(y_1, y_2, \ldots, y_k) = f_1(y_1) f_2(y_2 | y_1) \ldots f_k(y_k | y_1, \ldots, y_{k-1}) .$$

- Simulate each variable in turn to produce mutually consistent series

An approach based on generalised linear models

- **Idea**: model each variable in turn, in each case conditioning on previously-considered variables
  - **Justification**: generalised multiplication law — any joint distribution $f(y_1, y_2, \ldots, y_k)$ can be factorised as

$$f(y_1, y_2, \ldots, y_k) = f_1(y_1) f_2(y_2 | y_1) \ldots f_k(y_k | y_1, \ldots, y_{k-1}) \,.$$

- Simulate each variable in turn to produce mutually consistent series
- Component models for each variable are generalized linear models (GLMs):
  - Each value considered to be drawn from its own probability distribution
  - Distributions for each variable all of same form (normal, gamma, ...)
  - GLM-based WGs compete favourably with other state-of-the-art techniques (e.g. Maraun et al., *Rev. Geophys.*, 2010; Frost et al., *J. Hydrol.*, 2011)

GLMs for weather generation

- Means of distributions determined by linear functions of covariates representing, e.g., geographical location, time of year, indices of large-scale synoptic structure, previous days weather, current days values of other variables, . . .
- Variance usually determined by mean, but can be modelled separately
- Dependence on other variables ensures mutual consistency of generated series
  - **NB** dependence in one direction only (generalised multiplication law)
- Also need mutual consistency between spatial locations
  - Addressed using inter-site dependence models

Rglimclim

- Software package for developing multivariate, multisite daily weather generators using GLMs
- Runs under R (http://www.R-project.org) on all platforms
- Based on earlier Glimclim package — Fortran 77(!), multisite but univariate weather generator
- Adds graphical facilities and diagnostics as well as multivariate modelling / simulation capability
- Flexible model structures allow development based on physical understanding rather than statistical convenience
- Allows imputation of missing values (see later)

Modelling capability (I)

- Distributions currently available:
  - Normal (not very useful)
  - Heteroscedastic normal (suitable for, e.g., temperature)
  - Gamma (suitable for, e.g., wind speed, precipitation intensity)
  - Bernoulli (suitable for, e.g., precipitation occurrence)

Modelling capability (I)

- Distributions currently available:
  - Normal (not very useful)
  - Heteroscedastic normal (suitable for, e.g., temperature)
  - Gamma (suitable for, e.g., wind speed, precipitation intensity)
  - Bernoulli (suitable for, e.g., precipitation occurrence)
- Covariate classes:
  - 'Site effects': flexible representation of systematic regional variation ('climatology')
  - Seasonality: various options available
  - Autocorrelation: functions of lagged values
  - Inter-variable dependence: functions of simultaneous and lagged values of other variables
  - 'External' influences e.g. indices of large-scale climate
  - Interactions: allow effects of one variable to be modulated by others

Modelling capability (II)

- Several structures available for representing residual inter-site dependence to ensure spatial coherence
- Most based on correlation structures for standardised / Anscombe residuals (defined so as to have "almost Gaussian" distribution)

Modelling capability (II)

- Several structures available for representing residual inter-site dependence to ensure spatial coherence
- Most based on correlation structures for standardised / Anscombe residuals (defined so as to have "almost Gaussian" distribution)
- Additional options available for Bernoulli distributions — needed for realistic generation of spatial rainfall occurrence:
  - Thresholding of latent Gaussian field with spatial correlation structure — suitable for large regions
  - Beta-binomial representation for distribution of 'wet area' — suitable for small catchments where inter-site dependence is uniformly high
  - Model based on simple binary weather state process (original `Glimclim` model — other options preferable)

Model fitting and comparison

- Models fitted using maximum likelihood under (incorrect) assumption of independence between sites
  - Standard IWLS fitting algorithm, augmented to allow estimation of parameters in nonlinear covariate transformations
  - Computationally fast $\Rightarrow$ feasible to fit & compare many different models on large datasets
  - Lose some estimation efficiency compared with fully-specified spatial model — unimportant for large datasets
  - Usual standard errors adjusted for inter-site dependence (‘sandwich covariance estimation’)
- Model comparison using likelihood ratio tests adjusted for inter-site dependence (methodology of Chandler & Bate, *Biometrika*, 2007)
- Extensive summary and diagnostic information to identify lack-of-fit and guide model-building process

Simulation and imputation

- Simulated sequences can be either unconstrained (conventional WG) or conditioned on all available observations:
  - Allows for multiple imputation of missing observations ⇒ quantifies uncertainty in historical properties
  - Can also be used to 'interpolate' to regular grid — alternative to gridded datasets
- Summary and plot methods check ability to reproduce wide variety of properties

Example: the Thames again

Variables modelled and distributions used

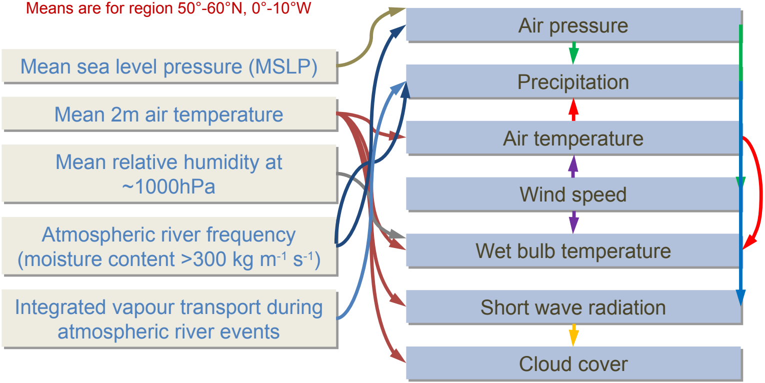
Variable	Distribution
Air pressure	Normal distribution with changing mean and variance
Rainfall	Logistic regression for occurrence (wet / dry), gamma distribution with changing mean & constant coefficient of variation (CV) for wet-day amounts
Air temperature	Normal distribution with changing mean and variance
Wind speed	Gamma distribution with changing mean & constant CV
Wet bulb temperature	Normal distribution with changing mean and variance
Short wave radiation	Gamma distribution with changing mean & constant CV
Cloud cover	Gamma distribution with changing mean & constant CV

Model fitted to data from 1970–2000

Thames: structure of multivariate model

Monthly indices of large-scale structure
Means are for region 50°-60°N, 0°-10°W

Local variables in weather generator



Thames: detail of windspeed model component

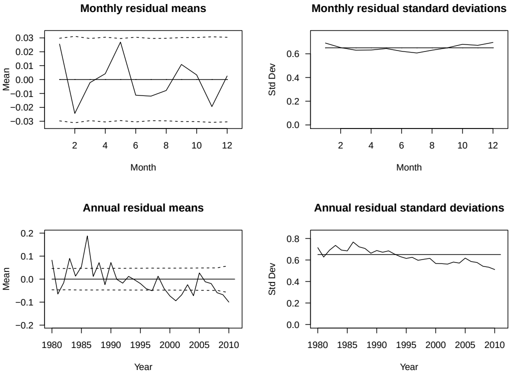
- Site s, day t: gamma distribution, mean μ_{st} & shape parameter α
  - $\log \mu_{st} = \beta_0 + \beta_1 x_{st}^{(1)} + \ldots + \beta_p x_{st}^{(p)}$ where $\{x_{st}^{(j)}\}$ are covariate values

```
> WindModel                               # name of stored wind speed model object
 WIND SPEED MODEL - GAMMA DISTRIBUTION
 =====================================

 Response variable: w_speed_ms

 Main effects:
 -------------
                                          Coefficient  Std Err   T-stat Pr(|T|>t)
                                 Constant     10.2918   0.9637  10.6800 < 2.2e-16
 1  Legendre polynomial  1 for Latitude        1.2369   0.1013  12.2084 < 2.2e-16
 2  Legendre polynomial  1 for Longitud       -0.3868   0.0288 -13.4091 < 2.2e-16
 ...
 7  Legendre polynomial  4 for Latitude       -0.5098   0.0255 -20.0141 < 2.2e-16
 8  Legendre polynomial  4 for Longitud       -0.1441   0.0034 -42.6402 < 2.2e-16
 9                         Mapped_Altitude     0.0331   0.0023  14.4920 < 2.2e-16
 10               altitude.std_dev_3x3km2     -0.1145   0.0121  -9.4966 < 2.2e-16
 11                          MSLP (code 51)     0.0574   0.0100   5.7106 1.127e-08
 12                         AR300 (code 54)     0.7323   0.0416  17.5900 < 2.2e-16
 13 Distance-weighted mean of air_pres[     -0.1535   0.0035 -43.6075 < 2.2e-16
 14 Daily seasonal effect, cosine compo      0.0491   0.0061   8.0249 1.018e-15
 15 Daily seasonal effect, sine compone      0.1146   0.0063  18.3253 < 2.2e-16
 ...
```

Checking the fit: seasonality and trends

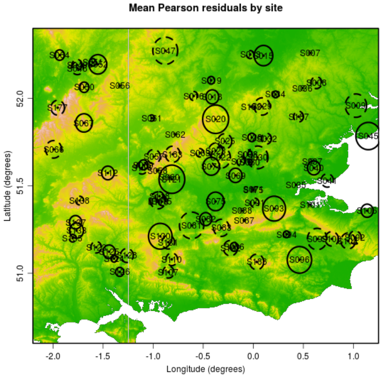


Code to generate these plots:

```
> par(mfrow=c(2,2)) # 2*2 array of plots
> plot(WindModel)
```

- Plots enable quick visualisation of unexplained structure in mean and variability
- Unexplained trend in annual means suggests model needs improving
  - Subsequent investigation revealed spurious trends in pressure covariate data

Checking the fit: systematic regional variation

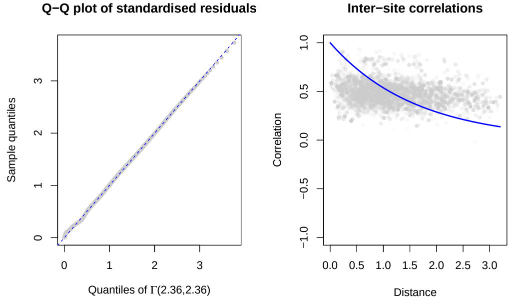


- Circle sizes proportional to average residual ("standardised model bias") at each site
- Solid & dashed lines indicate under- and overprediction
- Thick lines indicate residuals significantly different from zero (5% level)
- Some big residuals but no systematic structure — regional variation captured OK by model

Code to generate this plot:

```
> par(mfrow=c(1,1)) # Single plot on page
> plot(WindModel, which.plots=3,
  site.options=list(add.to.map=TRUE,scale=1.5,coord.cols=2:1))
```

Checking the fit: other diagnostics



Code to generate these plots:

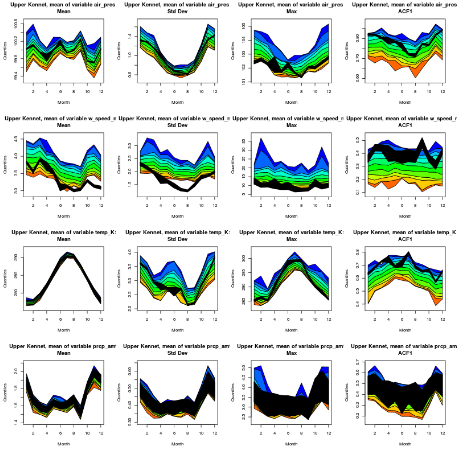
```
> par(mfrow=c(1,2)) # Two plots side by side
> plot(WindModel, which.plots=4:5,
       plot.cols=c(gray(0.8),"blue"), lwd=2)
```

- Quantile-quantile plot shows excellent fit of gamma distributions
- Observed inter-site residual correlations are all over the place . . .
  - **NB** shading intensity indicates # of observations contributing to each pairwise correlation — avoids overinterpreting very imprecise correlations

Thames: testing simulation performance

- 100 multivariate time series simulated simultaneously at station locations and Kennet grid nodes (357 locations total), 2001-2009 (validation period)
- Also 39 imputations
- Calculate summary measures for each simulation — 100 values for each summary
- Compare simulated distributions with envelope from imputations which is 95% interval for actual value

Thames: simulations (I)

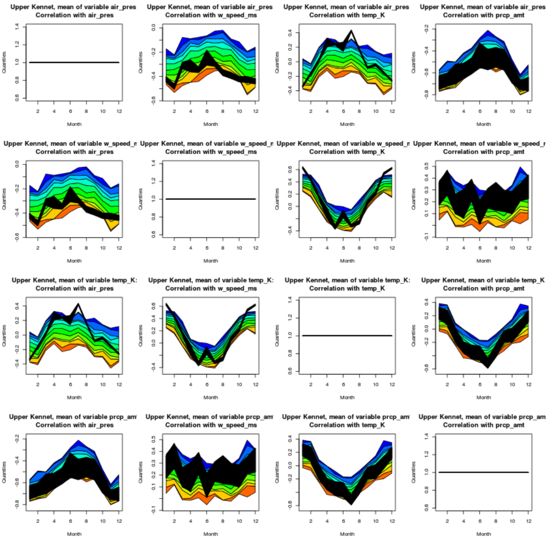


- Simulations overestimate wind speed variability here, otherwise OK
- Considerable uncertainty over precipitation due to lack of observations

Code to generate these plots:

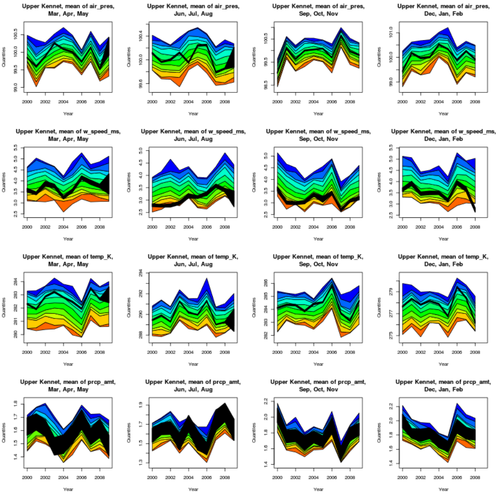
```
> par(mfrow=c(4,4))
> plot(sim.summary,
   imputation=obs.summary,
   which.timescales="daily",
   which.sites=NULL,
   which.regions="Upper Kennet",
   which.stats=c("Mean","Std Dev",
                 "Max","ACF1"),
   colours.sim="colour")
```

Thames: simulations (II)



- Same set of simulations, now looking at inter-variable correlations
- **NB** uncertainty over precipitation again

Thames: simulations (III)



- Now looking at seasonal means for each variable
- "Seasons" are user-defined
- Check for reproduction of interannual variability

Code to generate these plots:

```
> par(mfrow=c(4,4))
> plot(sim.summary,
   imputation=obs.summary,
   which.timescales="monthly",
   which.sites=NULL,
   which.regions="Upper Kennet",
   colours.sim="colour")
```

Concluding thoughts

- Package is powerful, flexible and computationally efficient compared with other *advanced* downscaling methods / weather generators
- Easy to produce diagnostics to assess suitability for use in impacts studies
- Provides information on uncertainty due to missing observations (cf gridded data products)

BUT

- Requires good level of statistical awareness — model-building not trivial

Obtaining the software:

Download from

http://www.homepages.ucl.ac.uk/~ucakarc/work/glimclim.html

Useful event?

3rd VALUE Training School: Spatial and Temporal Variability in Statistical and Dynamical Downscaling, Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste, Italy, 3–14 November 2014

http://www.value-cost.eu/node/1143

☺ *Thank you for your attention* ☺