Environmental data science

Slides at *www.lancs.ac.uk/∼jonathan*

Philip Jonathan

April 2018

Outline

Overview

- **Context**
- **Modelling the physical environment**
  - extremes
  - sensing
  - uncertainty analysis
  - working across disciplines
- **Reasons to be excited!**

Thanks

- Shell
- Durham and Lancaster
- n others, $n \gg 0$

Lay of the land

- Climate change
- Population growth
- Economic development
- Urbanisation / migration
- Increasing risk awareness / aversion (environmental, medical, litigation, insurance, . . .)

- Food: land and ocean use
- Water: supply, flood, erosion
- Air: pollution / waste
- Energy: renewables mix
- Security: Connectivity / privacy

Context

Digital acoustic sensing [Shell]. 10kHz sampling for each of n locations.

complexity, heterogeneity $\Leftrightarrow$ beauty, opportunity

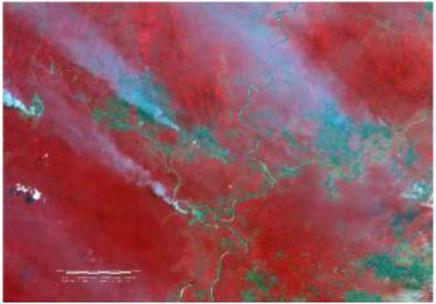
Accessible data

- $n_{2020} \gg n_{1980}$, $p_{2020} \gg p_{1980}$
- Streaming
- Connected data sources
- Numeric, text, images, sound, speech
- Increase in awareness of “data science”

Computing

- Parallelism: cores, clusters, GPU, memory, cloud
- Freeware: PYTHON, R, (C, JAVA)
- Data engineering: e.g. Alteryx, Spark, SQL

Context: models and uncertainty quantification



Smoke plume (Hirst et al. 2013)

More science

- Multi-scale
- ODEs, PDEs, dynamics
- Likelihood, extremes

More Bayesian

- Awareness, acceptance, interpretation
- Emulation, Gaussian processes
- Graphical models, dynamic linear models
- “Approximate” Bayesian methods
- Optimal decisions

Context: better, faster measurement and coverage

An ocean drifter [diydrones.com]
Diameter $\approx$ 20cm, 1000s deployed.

- Good, cheap, widget sensors
  - In-situ, bio-tracking, drifters, floaters
- Satellites
  - Ocean, seismic, GHGs, land use, telemetry
- Drones, autonomous vehicles, high-altitude pseudo-satellites
- Spectroscopy, optics, hyperspectral
- IoT

Context: connectivity

- Everything and everyone digitally inter-connected
- Everything and everyone feasible source data for empirical inference
- ... whether we like it or not

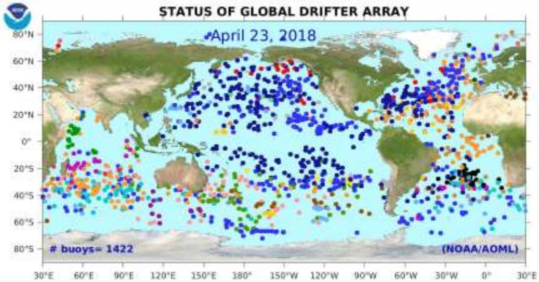
- Global infrastructure
- 10^n transactions per second, $n \uparrow$
- **New state for humanity?**

- "Crude" data "ingested" into "unstructured data store", subsequently "refined" and extracted to structured data "data mart" or "data lake"
- Inference on data mart using "analytics"

[Microsoft]

Context: connectivity

EO satellite coverage, 24.04.18 22.15pm [in-the-sky.org]. Only weather, NOAA, GOES, Earth Resources, SARSAT, Disaster Monitoring, Tracking and Data Relay Satellites, ARGOS, Planet, Spire shown.



Drifter coverage, [NOAA]

- Waves: 9 altimeters, 12 radiometers, 3 scatterometers [1980-2014; Young 2016]
- CH_4: Sentinel 5G / Tropomi [2017-date; ESA]
- ≈ 1500 drifters measuring temperature, surface current, dispersion of surface particles
- Computing resources: e.g. JASMIN [CEDA]

Modelling the physical environment

Spatio-temporal extremes

- Marine environment: wave, wind and current fields. Short- and long-term hazards
- Planetary and atmospheric-oceanic interactions, different processes, scales
- Measurements (altimetry, radar, laser, buoy)
- Complex physical models (e.g. genesis-track)

- Rich asymptotic theory: extreme value analysis
- Spatio-temporal, non-stationary, multivariate
- Typical sparse data (tails), multi-source

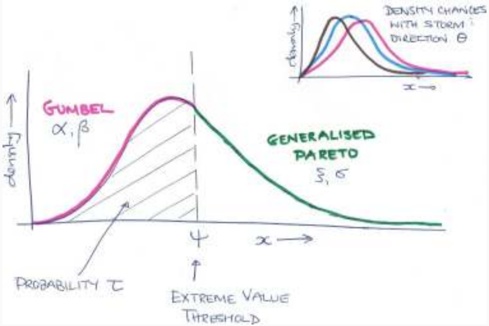
- Heatwave, drought, earthquake, solar flare, . . .

Roker lighthouse, Sunderland [Daily Express].

Huge scope and requirement for research in almost all aspects
including **non-stationarity** and **uncertainty quantification** in particular

Non-stationary marginal extremes: gamma-GP model

- Sample of peaks over threshold y, with covariates θ
  - θ is 1D (directional) here, could be nD (space, time, direction, season, ...)
- Below threshold ψ
  - $y \sim$ truncated gamma with shape α, scale $1/\beta$
- Above ψ
  - $y \sim$ generalised Pareto with shape ξ, scale σ
- ξ, σ, α, β, ψ all functions of θ
- $\Pr(X < \psi|\theta) = \tau$
- Likelihood here



A gamma-generalised Pareto model (Randell et al. 2016)

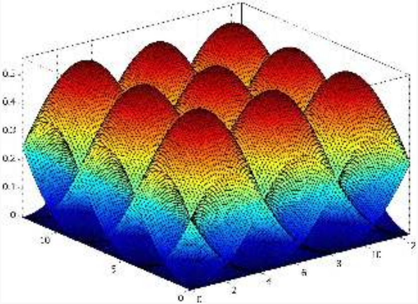
Covariate effects critical but intricate $\Leftrightarrow$ algorithms, computation

Non-stationary marginal extremes: P-splines

- Physics: $\alpha, \beta, \rho, \xi, \sigma, \psi$ vary smoothly with θ
- B-spline basis $\boldsymbol{B}$ on index set of covariates
- For $\eta \in \{\alpha, \beta, \rho, \xi, \sigma, \psi\}$, write $\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\beta}_\eta$
- In nD, $\boldsymbol{B} = \boldsymbol{B}_{\theta_n} \otimes \ldots \otimes \boldsymbol{B}_{\theta_\kappa} \otimes \ldots \otimes \boldsymbol{B}_{\theta_2} \otimes \boldsymbol{B}_{\theta_1}$

- Spline roughness for dimension $\kappa \sim \lambda_{\eta\kappa} \beta'_{\eta\kappa} \boldsymbol{P}_{\eta\kappa} \boldsymbol{\beta}_{\eta\kappa}$
- Penalty $\boldsymbol{P}_{\eta\kappa}$ function of stochastic roughnesses $\delta_{\eta\kappa}$

- B-splines local support, GLAMs for slick computation



Kronecker product of marginal spline bases.

Scope for more scalable descriptions, algorithms in nD $\Leftrightarrow$ adaptive splines, reweighed kernels

Non-stationary marginal extremes: priors, conditional structure and algorithms

Priors

$$\begin{aligned} \text{density of } \boldsymbol{\beta}_{\eta\kappa} &\propto \exp\left(-\frac{1}{2}\lambda_{\eta\kappa}\boldsymbol{\beta}'_{\eta\kappa}\boldsymbol{P}_{\eta\kappa}\boldsymbol{\beta}_{\eta\kappa}\right) \\ \lambda_{\eta\kappa} &\sim \text{gamma} \\ \tau &\sim \text{beta} \end{aligned}$$

Full conditionals for $\Omega = \{\alpha, \beta, \rho, \xi, \sigma, \psi, \tau\}$

$$\begin{aligned} f(\tau|\boldsymbol{y}, \Omega \setminus \tau) &\propto f(\boldsymbol{y}|\tau, \Omega \setminus \tau) \times f(\tau) \\ f(\boldsymbol{\beta}_\eta|\boldsymbol{y}, \Omega \setminus \boldsymbol{\beta}_\eta) &\propto f(\boldsymbol{y}|\boldsymbol{\beta}_\eta, \Omega \setminus \boldsymbol{\beta}_\eta) \times f(\boldsymbol{\beta}_\eta|\boldsymbol{\delta}_\eta, \boldsymbol{\lambda}_\eta) \\ f(\boldsymbol{\lambda}_\eta|\boldsymbol{y}, \Omega \setminus \boldsymbol{\lambda}_\eta) &\propto f(\boldsymbol{\beta}_\eta|\boldsymbol{\delta}_\eta, \boldsymbol{\lambda}_\eta) \times f(\boldsymbol{\lambda}_\eta) \end{aligned}$$

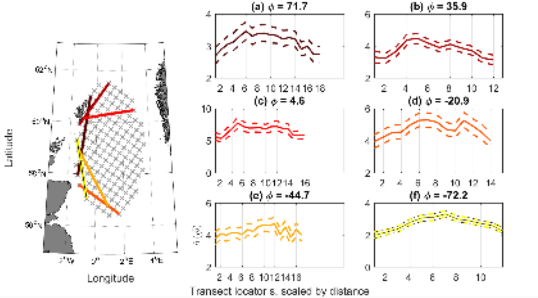
Problem size

- $p \approx 5 \times 10^3$ for $\theta\phi$, and $\approx 3 \times 10^7$ for $XY\theta\phi$
- HPC, `MATLAB` cluster

Algorithms

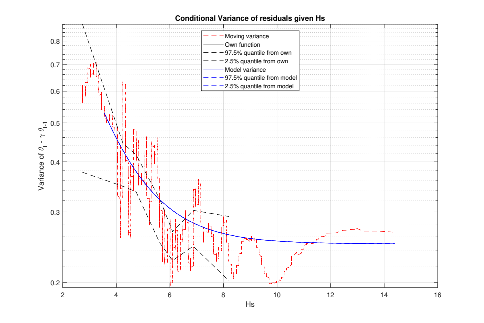
- Elements of $\boldsymbol{\beta}_\eta$ highly interdependent, correlated proposals essential for good mixing
- "Stochastic analogues" of IRLS and back-fitting algorithms
- Estimation of different penalty coefficients for each covariate dimension
- Gibbs sampling when full conditionals available
- Otherwise Metropolis-Hastings (MH) within Gibbs, using suitable proposal mechanisms including adaptive MCMC and mMALA where possible

Spatial extremes



Spatial dependence (Ross et al. 2017)

Conditional extremes

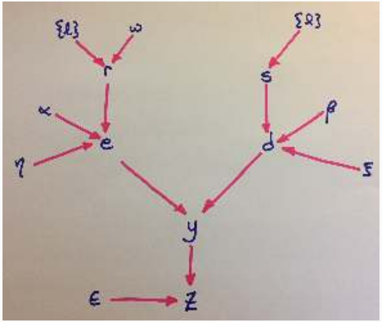


Directional variance with H_S (Randell et al. 2018)

- Extreme ocean storms
- Max-stable process
- Non-stationary extremal dependence
- Maths here

- Storm evolution in time and direction
- Non-stationary Markov extremal model
- Dynamic model for direction
- Maths here

Extremes: scope and requirement for research in almost all aspects



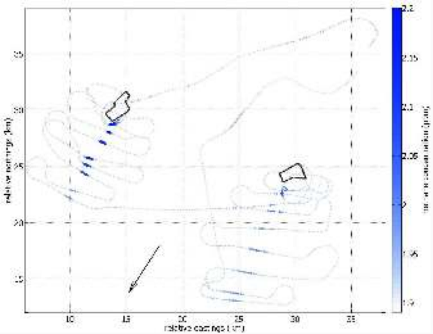
A simple system model

- Flexible framework, Bayes linear
- Optimal design (Jones et al. 2015, 2018a)
- Extreme environments (Jones et al. 2018b)
- Probabilistic ODEs, Bayesian optimisation

Obs	:	$z(x) = y(x) + \epsilon$
Sys	:	$y(x) = e(x) + d(x)$
Emul	:	$e(x) = \alpha' g(x) + r(x, \omega) + \eta$
Disc	:	$d(x) = \beta' h(x) + s(x) + \xi$
e	:	emulator or "process" model
d	:	discrepancy model
g, h	:	non-linear bases for covariate space
r, s	:	Gaussian process residuals
Priors	:	all Gaussian
Data	:	emulator E, measured Z
Estimation	:	$f(\alpha, \beta, \{\ell_r\}, \{\ell_s\}, \omega \vert E, Z)$
Prediction	:	$f(y(x) \vert E, Z)$

General purpose, scalable approach to quantify system uncertainty

Probabilistic inversion



Airborne sensing (Hirst et al. 2013)

Airborne and line-of-sight sensing (Hirst et al. 2013, 2017)

- Trace concentrations (ppb) of gases, particulates
- Transported on wind from source
- Sensitive optical point or line-of-sight sensors

- Wind field known approximately
- Background can be problematic (CO_2)
- Measurement error

Probabilistic inversion

Model: $y = A(\alpha, \delta_\phi)s(\{z, w, \rho\}) + b(\beta) + \epsilon(\sigma)$

Physics

- Sources s: multiple, spiky; Gaussian mixture
- Background b: smooth; Gaussian Markov random field, wind covariate
- Plume A: Gaussian

Parameters

- Source locations z, "widths" w and emission rates ρ for mixture of m sources
- Random field background parameters β
- Measurement error standard deviation σ_ϵ
- Wind–direction correction δ_ϕ
- Others (e.g. plume opening angles α)

Set-up

- Static: point, line-of-sight
- Dynamic: vehicular, airborne

Inference

- Reversible jump MCMC inference over sources

Opportunities

- Multiple responses
- Forward model
- Non-stationary sources
- Design of measurement campaigns
- Other processes

Working across disciplines: statistical know-how

Experimental process

- Design
- Measurement
- Exploration, visualisation
- Estimation
- Prediction, detection
- Validation
- Deployment

System design

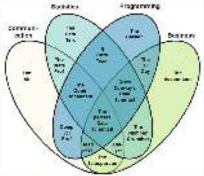
- Data engineering
- Software design

Methods

- **Communication and consultancy**
- DoE: factorial, CC, space-filling
- Sampling
- Data reduction: PCA, clustering
- Regularisation: ridge, LASSO, elastic net
- Non-parametric: Gaussian processes, trees
- Model selection, evidence
- Model checking: cross-validation, bootstrapping, randomised permutation testing
- Vanilla MH / MCMC
- . . .

Skill set

[Drew Conway]



[Yanir Seroussi]

Excellence in inter-disciplinary research requires fit-for-purpose statistical thinking and modelling

Reasons to be excited!

Large-scale environmental inference (more here)

- Non-stationarity
- Bayesian uncertainty analysis
- Scalability
- Complexity (e.g. solitons, plume evolution, overturning circulation)

Successful inter-disciplinary research

- Pragmatism, parsimony, impact
- Tailored solutions (e.g. exploit sparsity in typhoon modelling)

Physical environment

- Rich physics: multi-scale, dynamics
- Measurement: multi-source, multi-type
- "Data fusion" and calibration
- Global societal impact

Statistical inference

- Sample scale, size and speed
- Non-stationary, spatio-temporal
- Multivariate
- Likelihoods informed by theory and physics
- Bayesian inference, emulation
- Uncertainty quantification, optimal decisions

Statistical practice

- Good statistical thinking, parsimony
- Ethics, responsibility, accountability

Computation

- Slick algorithms exploiting architecture
- System design, data engineering

Useful integrated physical, measurement, computational and statistical science

References

B. Hirst, P. Jonathan, F. González del Cueto, D. Randell, and O. Kosut. Locating and quantifying gas emission sources using remotely obtained concentration data. *Atmospheric Environ.*, 74:141–158, 2013.

B. Hirst, D. Randell, M. Jones, P. Jonathan, B. King, and M. Dean. A new technique for monitoring the atmosphere above onshore carbon storage projects that can estimate the locations and mass emission rates of detected sources. *Energy Procedia*, 114:3716–3728, 2017.

P. Jonathan, K. C. Ewans, and D. Randell. Non-stationary conditional extremes of northern North Sea storm characteristics. *Environmetrics*, 25:172–188, 2014.

M. J. Jones, M. Goldstein, P. Jonathan, and D. Randell. Bayes linear analysis for Bayesian optimal experimental design. *J. Stat. Plan. Inference*, 171:115–129, 2015.

M. J. Jones, M. Goldstein, P. Jonathan, and D. Randell. Bayes linear analysis of sequential optimal design problems. *(Submitted to the Electron. J. Stat. in January 2018, draft at www.lancs.ac.uk/∼jonathan)*, 2018a.

M. J. Jones, H. F. Hansen, A. R. Zeeberg, D. Randell, and P. Jonathan. Uncertainty quantification in estimation of ocean environmental return values. *(Submitted to Coastal Eng. in February 2018, draft at www.lancs.ac.uk/∼jonathan)*, 2018b.

D. Randell, K. Turnbull, K. Ewans, and P. Jonathan. Bayesian inference for non-stationary marginal extremes. *Environmetrics*, 27:439–450, 2016.

D. Randell, S. Tendijk, E. Ross, and P. Jonathan. A statistical model for the evolution of extreme storm events. *(In preparation for Environmetrics, April 2018)*, 2018.

E Ross, M Kereszturi, M van Nee, D Randell, and P Jonathan. On the spatial dependence of extreme ocean storm seas. *Ocean Eng.*, 145:1–14, 2017.

R. Towe, E. Eastoe, J. Tawn, and P. Jonathan. Statistical downscaling for future extreme wave heights in the North Sea. *Ann. Appl. Stat.*, 11:2375–2403, 2017.

Non-stationary marginal extremes: simple gamma-GP model (back)

- Density is $f(y|\xi,\sigma,\alpha,\beta,\psi,\tau)$

$$= \begin{cases} \tau \times f_{TG}(y|\alpha,\beta,\psi) & \text{for } y \leq \psi \\ (1-\tau) \times f_{GP}(y|\xi,\sigma,\psi) & \text{for } y > \psi \end{cases}$$

- Likelihood is $\mathcal{L}(\xi,\sigma,\alpha,\beta,\psi,\tau|\{y_i\}_{i=1}^n)$

$$= \prod_{i:y_i \leq \psi} f_{TG}(y_i|\alpha,\beta,\psi) \prod_{i:y_i > \psi} f_{GP}(y_i|\xi,\sigma,\psi)$$
$$\times \quad \tau^{n_B}(1-\tau)^{(1-n_B)} \text{ where } n_B = \sum_{i:y_i \leq \psi} 1$$

- Estimate all parameters as functions of θ

- Locations $\{s_k\}_{k=1}^p$, maxima $\{X_k\}$, covariates $\{\mathcal{C}_k\}$, density $\dot{f}$, cdf $\dot{F}$
- $\dot{f}(x_1, x_2, ..., x_p) = \left[\dot{f}(x_1)\dot{f}(x_2)...\dot{f}(x_p)\right] \dot{f}(x_1, x_2, ..., x_p)$
- $X_k \sim \text{GEV}(\xi_k, \beta_k, \mu_k)$, so $\dot{f}, \dot{F}$ known
- GEV parameters ξ_k, β_k, μ_k vary smoothly between locations, and with $\mathcal{C}_k$
- Frechet scale: $x \to z$; $\dot{f}, \dot{F} \to f, F$
- $F(z_1, z_2, ..., z_p) = \exp\{-V(z_1, z_2, ..., z_p)\}$
- $V_{kl}(z_k, z_l; h(\Sigma)) = \frac{1}{z_k}\Phi(\frac{m(h)}{2} + \frac{\log(z_l/z_k)}{m(h)}) + \frac{1}{z_l}\Phi(\frac{m(h)}{2} + \frac{\log(z_k/z_l)}{m(h)})$
- $h = s_l - s_k$, $m(h) = (h'\Sigma^{-1}h)^{1/2}$, Φ is Gaussian
- Covariate effects $\mathcal{C}$ in Σ
- Joint Bayesian inference for $\{\xi_k(\mathcal{C}), \sigma_k(\mathcal{C}), \mu_k(\mathcal{C})\}$ and $\Sigma(\mathcal{C})$

Basics

- Y_1, Y_2 on standard Laplace scale via non-stationary marginal modelling
- For large class of joint distributions, we have $(Y_2|Y_1 = y_1) = \alpha_{21} y_1 + y_1^{\beta_{21}} W_{21}$ for $y_1 > \phi_1$,
- ϕ_1 large, $\alpha \in [-1, 1]$, $\beta \in (-\infty, 1]$
- W_{21} estimated from regression residuals
- Easily extended to p dimensions with non-stationarity

(Heffernan and Tawn 2004, Jonathan et al. 2014)

Storm evolution

- $\{Y_t, \theta_t\}$, $Y_t \sim$ Laplace, $\theta_t \in [0, 2\pi)$
- $\mathtt{MEM}(\tau)$ for order τ: $(Y_{t+\tau}|Y_t = y) = \alpha_\tau y + y^{\beta_\tau} W_{t+\tau|t,t+1,\ldots,t+\tau-1}$
- W estimated by kernels
- Direction: $\Delta\theta_{t+1} = \gamma_1 \Delta\theta_t + \gamma_2 \Delta\theta_{t-1} + \epsilon_t$
- $\text{var}\epsilon_t = f(Y_t^o)$

(Winter and Tawn 2015, Randell et al. 2018)

UQ

- Hierarchical model
  - Stephenson (2009)
  - Reich and Shaby (2012)
  - Allows DLM, emulation, UQ
  - Asymmmetric logistic dependence (so AD)
- Statistical downscaling
  - e.g. Towe et al. [2017]
- Non-stationarity
  - Arbitrary covariate representations

Multi-source

- Physical model ("hindcast") is basic framework
- Complementary measurements (e.g. satellites)
- HT calibration
- Non-stationarity
- Vanilla version in Jones et al. [2018b]