Interactive Visual Analysis of Multi-Dimensional Scientific Data

Helwig Hauser (et al.),
Dagstuhl, **2011-06-06**

Scientific data – one common view

- **Measured or simulated** data of some phenomenon
- Usually in relation to **space** and/or **time**
- Useful data model **d**(**x**) with
  - domain **x** (usually 2D or 3D space, time), independent variables
  - range **d** (measured/simulated values), dependent values (dependent on **x**)
- Typical examples:
  - CT data $d(\mathbf{x})$ with $\mathbf{x} \in R^3$ and $d \in R$
  - unsteady 2D flow $\mathbf{v}(\mathbf{x},t)$ with $\mathbf{x} \in R^2$, $t \in R$, and $\mathbf{v} \in R^2$
  - flow sim. result $\mathbf{d}(\mathbf{x},t)$ with $\mathbf{x} \in R^3$, $t \in R$, and $\mathbf{d} \in R^n$
- Dimensionality?

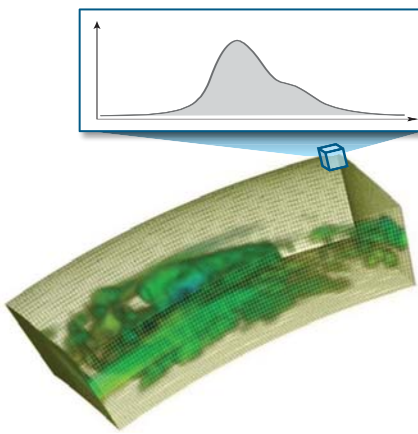
High-dimensional vs. multi-variate data

- **Increased dimensionality** of **d**(**x**) wrt.
  - range **d** → multi-variate data
  - domain **x** → multi-dimensional data
- Very often in SciVis: neither **d** nor **x** is high-dim.! Examples: CT scan, vector field, etc.
- Also addressed: **multi-variate scientific data**
  - multi-variate simulation data
    - ex.: simulated Diesel particulate filter, $\mathbf{x}\in R^3\times R$, $\mathbf{d}\in R^{37}$: range 37-dim. (or so)
    - integrated visualization, IVA w/ L&B, dim.-reduction techniques, etc.
  - multi-modal measurements, …

- Multi-dimensional scientific data?

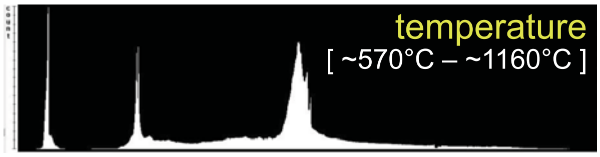
Multi-dimensional scientific data

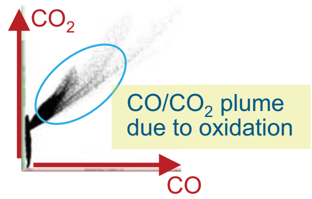
- More independent dimensions (more than space & time)
- One interesting & challenging class: **multi-run / ensemble** data
  - set of datasets, f.i.,
    - perturbed physics ensemble
    - initial condition ensemble
  - data $\mathbf{d}(\mathbf{s},t,\mathbf{p})$ w/ $\mathbf{d}\in R^n$ – can be multi-variate, too – dependent on
    - space **s** (2D or 3D)
    - time *t* (or not)
    - parameter(s) **p**, $\mathbf{p}\in R^m$
  - dealt with in
    - climatology
    - engineering
    - …

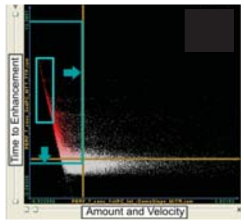


Interactive visual analysis

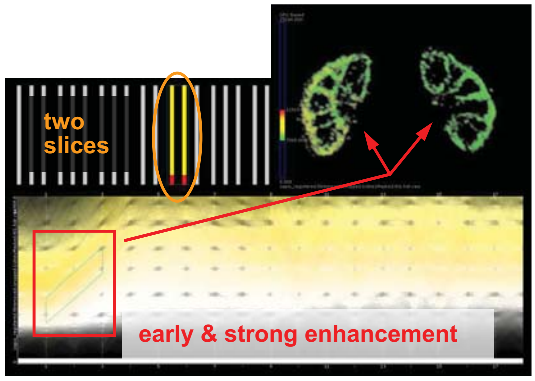
- Understanding data **wrt. range d**
  - which distribution has data attribute $\mathbf{d}_i$?
  - how do $\mathbf{d}_i$ and $\mathbf{d}_j$ relate to each other?
  - which $\mathbf{d}_k$ discriminate data features?





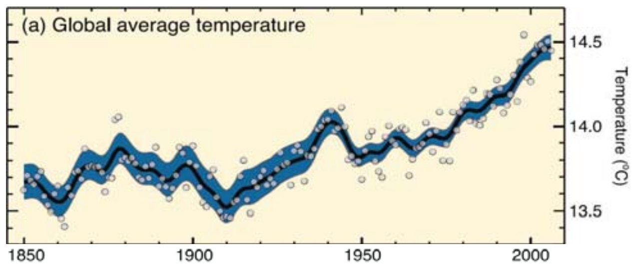


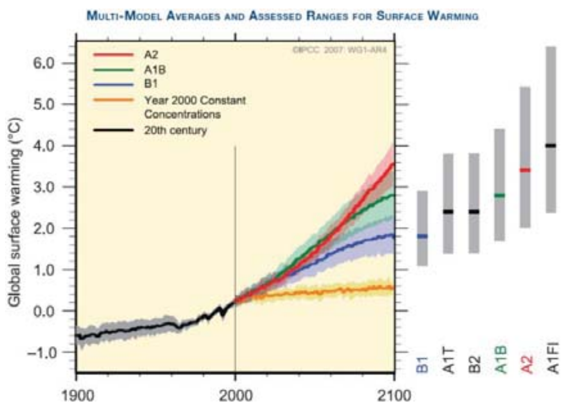
- Understanding data **wrt. domain x** (**s**, *t*, and **p**)
  - **where** (in **s**) are interesting data features?
  - **when** (in *t*) do they happen?
  - how are they **related to parameters p**?

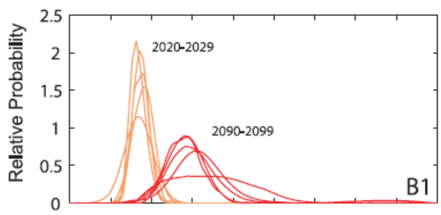


Investigating multi-run / ensemble data

- Often: visualizing statistics
  - trend (e.g., mean) & variation (std. dev.)
  - data quartiles (e.g., via boxplots)
- Also: comparing aggregates
  - statistics per run (class of runs)
  - overlay of aggregates per run
- One goal: sensitivity analysis





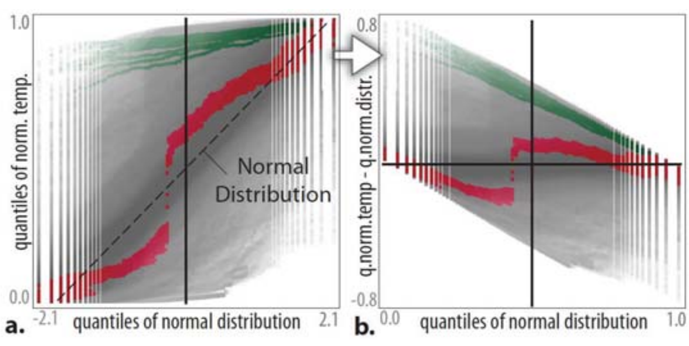


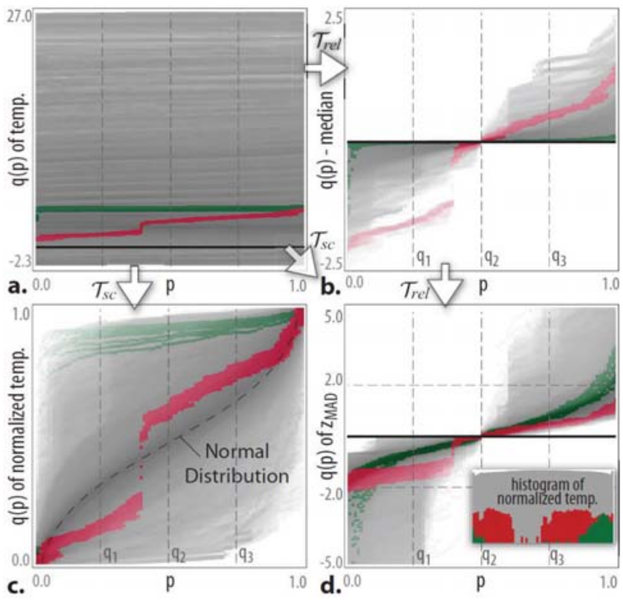
Trends vs. outliers wrt. parameter space

- Statistical aggregates across **p** to analyze
  - **trends**
    - mean μ, std.-dev. σ, …
    - quartiles q_1, q_2 (median), and q_3, IQR $q_3 - q_1$, …
    - octiles $e_i = q(i/8)$, quantiles $q(p)$ with $p \in [0,1]$
    - …
  - **outliers**
    - mild outliers: $< q_1 - 1.5 \cdot \mathrm{IQR}$, $> q_3 + 1.5 \cdot \mathrm{IQR}$
    - strong outliers: $< q_1 - 3 \cdot \mathrm{IQR}$, $> q_3 + 3 \cdot \mathrm{IQR}$
    - data outside $[-2,2]$ after z-standardization: $z = (x_i - \mu)/\sigma$
    - …
- Computing **multi-variate** (statistics) data (per $\mathbf{s} \times t$) **from multi-dimensional** (raw) data

Visualizing / analyzing lots of statistics

- Through statistical aggregation:
  - multiple statistics per spatiotemporal location
  - from $d(\mathbf{s},t,\mathbf{p})$ to $\mu(\mathbf{s},t)$, $\sigma(\mathbf{s},t)$, etc.
- Useful views allow the interactive visual analysis
  - **quantile-plot** $q(p)$ vs. p, here for numerous $(\mathbf{s},t)$
  - **detrending** (e.g., $-q_2$), **normalization** (e.g., z)



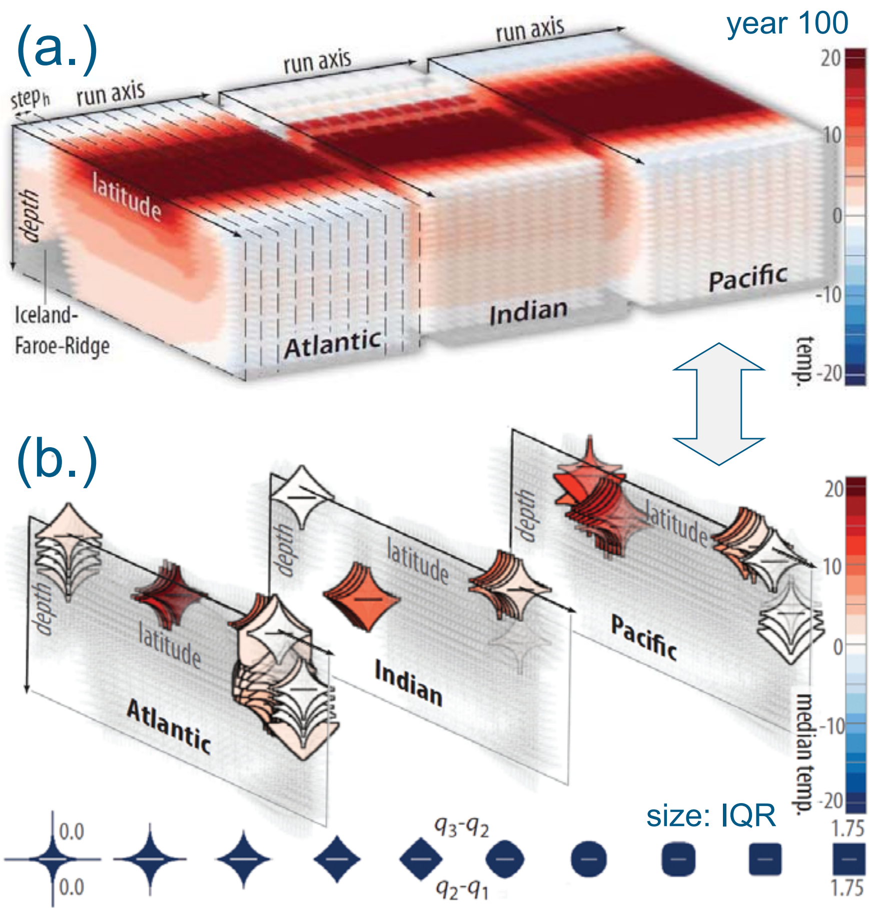


Linking ensemble data and aggregates

- Climate sim. data (temp.)
- Independent dimensions:
  - 3* 2D slices (lat.×depth)
  - 500 years
  - 2 params. (10×10)

a. All runs along 3rd dimension

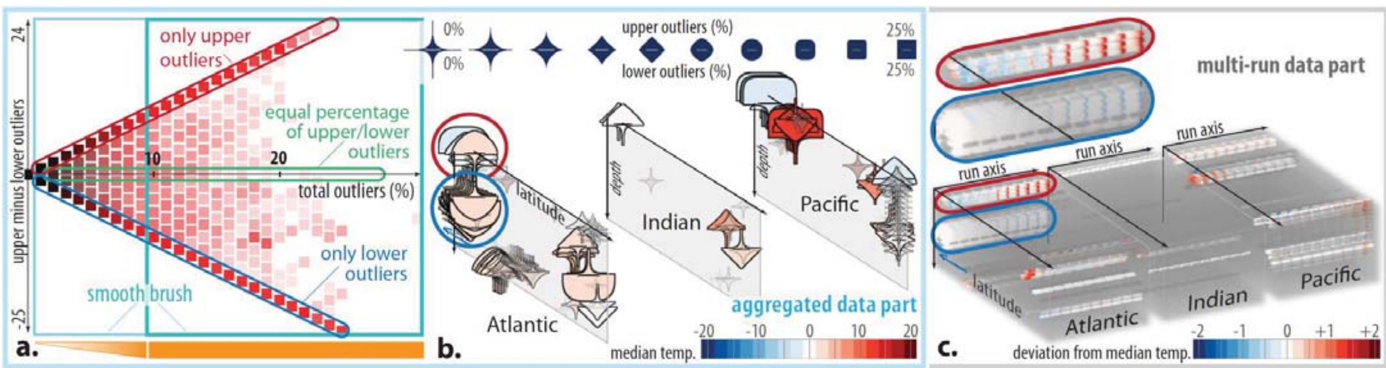
b. Glyphs show temp.-stats.



Relating aggregates and raw data

a. %outliers vs. upper–lower outliers (≥10%outliers brushed)

b. linked glyphs locate brushed locations



c. linked raw data vis shows responsible outlier runs

Discussion, conclusions, questions

- **Multi-field** can be **multi-variate** or **multi-dimensional**
- **Multi-dimensional** scientific data **challenging, integrated statistical aggregation** can help
- **Visualization** on at least **two** (linked) **levels**
  - **aggregates** (transforming dims. into variates)
  - **original** (multi-dim.) **data**
- IVA reveals **trends**, but also **outliers,** helps to understand **distributions**
- Understanding **data features from statistical aggregates** is **challenging**
- Up to now according to an **Eulerian perspective** – Lagrangian, feature-based perspective?
- Other forms of multi-dimensional scientific data?

Acknowledgements

- **Johannes Kehrer** ( ), **Peter Filzmoser** (TU WIEN), et al.
- **PIK** (Potsdam), **SimVis** (Vienna)
- **You! :–)**

- **Brushing Moments in Interactive Visual Analysis** by J. Kehrer et al.; CGF 29(3):813–822, 2010
- **Interactive Visual Analysis of Heterogeneous Scientific Data across an Interface** by J. Kehrer et al.; TVCG 17(7):934–946, 2011