# From open source and open data to “open computation”

## A climate science perspective

Roly Perera
Institute of Computing for Climate Science, University of Cambridge
School of Computer Science, University of Bristol

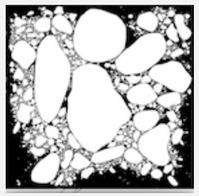
# ICCS overview

director + 3 co-directors
programme manager
3 researchers
9 RSEs

Institute of Computing for Climate Science

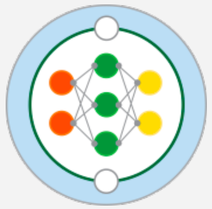
Collaboration/RSE support for 4 VESRIs (Virtual Earth Systems Research Institutes)

DataWave: Collaborative Gravity Wave Research

SASIP: Scale-Aware Sea Ice Project

LEMONTREE: Land Ecosystem Models based On New Theory, Observations, and Experiments

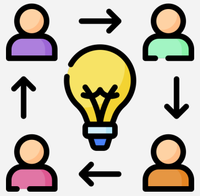


M²LInES: Multiscale Machine Learning In Coupled Earth System Modeling

# Ingredients of open science

What is open science?

open source



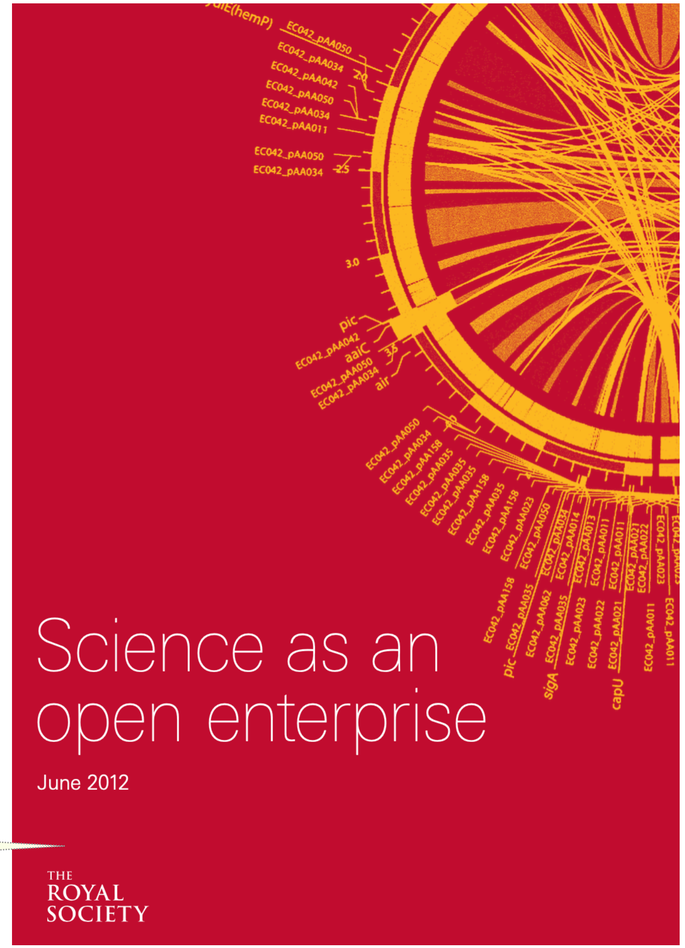
working in the open

open data

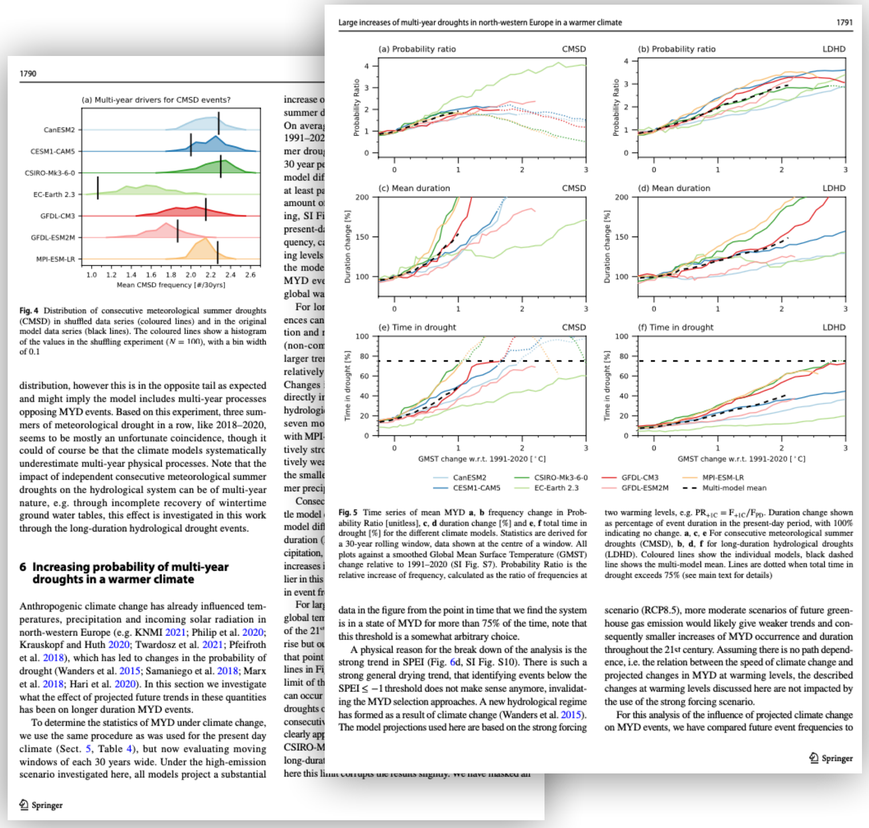
But what else?

Not a magic sauce that makes outcomes transparent and intelligible

Complex relationships between data and outcomes need to be **accessible**, **intelligible**, and easy to **evaluate** and **reuse**



# Open science UX today



1790

(a) Multi-year drivers for CMSD events?

CanESM2
CESM1-CAM5
CSIRO-Mk3-6-0
EC-Earth 2.3
GFDL-CM3
GFDL-ESM2M
MPI-ESM-LR

1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6
Mean CMSD frequency [#/30yrs]

**Fig. 4** Distribution of consecutive meteorological summer droughts (CMSD) in shuffled data series (coloured lines) and in the original model data series (black lines). The coloured lines show a histogram of the values in the shuffling experiment ($N = 100$), with a bin width of 0.1

distribution, however this is in the opposite tail as expected and might imply the model includes multi-year processes opposing MYD events. Based on this experiment, three summers of meteorological drought in a row, like 2018–2020, seems to be mostly an unfortunate coincidence, though it could of course be that the climate models systematically underestimate multi-year physical processes. Note that the impact of independent consecutive meteorological summer droughts on the hydrological system can be of multi-year nature, e.g. through incomplete recovery of wintertime ground water tables, this effect is investigated in this work through the long-duration hydrological drought events.

## 6 Increasing probability of multi-year droughts in a warmer climate

Anthropogenic climate change has already influenced temperatures, precipitation and incoming solar radiation in north-western Europe (e.g. KNMI 2021; Philip et al. 2020; Krauskopf and Huth 2020; Twardosz et al. 2021; Pfeifroth et al. 2018), which has led to changes in the probability of drought (Wanders et al. 2015; Samaniego et al. 2018; Marx et al. 2018; Hari et al. 2020). In this section we investigate what the effect of projected future trends in these quantities has been on longer duration MYD events.

To determine the statistics of MYD under climate change, we use the same procedure as was used for the present day climate (Sect. 5, Table 4), but now evaluating moving windows of each 30 years wide. Under the high-emission scenario investigated here, all models project a substantial

Springer

Large increases of multi-year droughts in north-western Europe in a warmer climate 1791

(a) Probability ratio CMSD
(b) Probability ratio LDHD
(c) Mean duration CMSD
(d) Mean duration LDHD
(e) Time in drought CMSD
(f) Time in drought LDHD

Probability Ratio
Duration change [%]
Time in drought [%]
GMST change w.r.t. 1991-2020 [ °C]

CanESM2 CSIRO-Mk3-6-0 GFDL-CM3 MPI-ESM-LR
CESM1-CAM5 EC-Earth 2.3 GFDL-ESM2M Multi-model mean

**Fig. 5** Time series of mean MYD **a**, **b** frequency change in Probability Ratio [unitless], **c**, **d** duration change [%] and **e**, **f** total time in drought [%] for the different climate models. Statistics are derived for a 30-year rolling window, data shown at the centre of a window. All plots against a smoothed Global Mean Surface Temperature (GMST) change relative to 1991–2020 (SI Fig. S7). Probability Ratio is the relative increase of frequency, calculated as the ratio of frequencies at two warming levels, e.g. $PR_{+1C} = F_{+1C}/F_{PD}$. Duration change shown as percentage of event duration in the present-day period, with 100% indicating no change. **a**, **c**, **e** For consecutive meteorological summer droughts (CMSD), **b**, **d**, **f** for long-duration hydrological droughts (LDHD). Coloured lines show the individual models, black dashed line shows the multi-model mean. Lines are dotted when total time in drought exceeds 75% (see main text for details)

data in the figure from the point in time that we find the system is in a state of MYD for more than 75% of the time, note that this threshold is a somewhat arbitrary choice.

A physical reason for the break down of the analysis is the strong trend in SPEI (Fig. 6d, SI Fig. S10). There is such a strong general drying trend, that identifying events below the SPEI $\leq -1$ threshold does not make sense anymore, invalidating the MYD selection approaches. A new hydrological regime has formed as a result of climate change (Wanders et al. 2015). The model projections used here are based on the strong forcing scenario (RCP8.5), more moderate scenarios of future greenhouse gas emission would likely give weaker trends and consequently smaller increases of MYD occurrence and duration throughout the 21st century. Assuming there is no path dependence, i.e. the relation between the speed of climate change and projected changes in MYD at warming levels, the described changes at warming levels discussed here are not impacted by the use of the strong forcing scenario.

For this analysis of the influence of projected climate change on MYD events, we have compared future event frequencies to

Springer

zenodo Search Upload Communities

June 29, 2022 Software Open Access

Scripts for "Large increases of multi-year droughts in north-western Europe in a warmer climate"

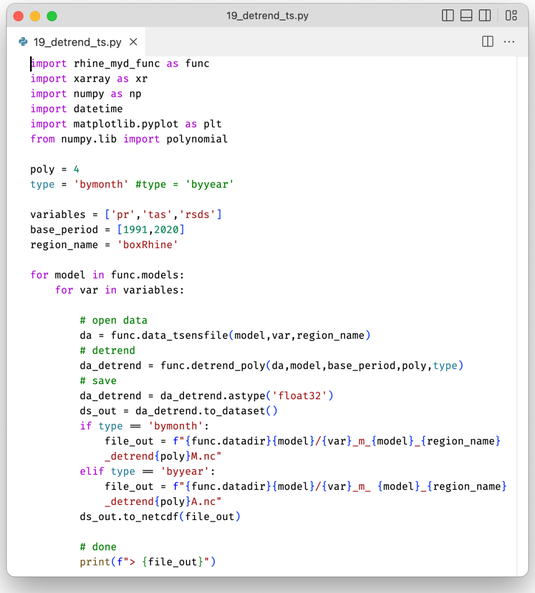
Van der Wiel, Karin; Batelaan, Thomas J; Wanders, Niko

Scripts associated with **Van der Wiel, Batelaan, Wanders (202x): Large increases of multi-year droughts in north-western Europe in a warmer climate,** *Climate Dynamics,* **DOI** 10.1007/s00382-022-06373-3.

Please read the README file for more information on the scripts.

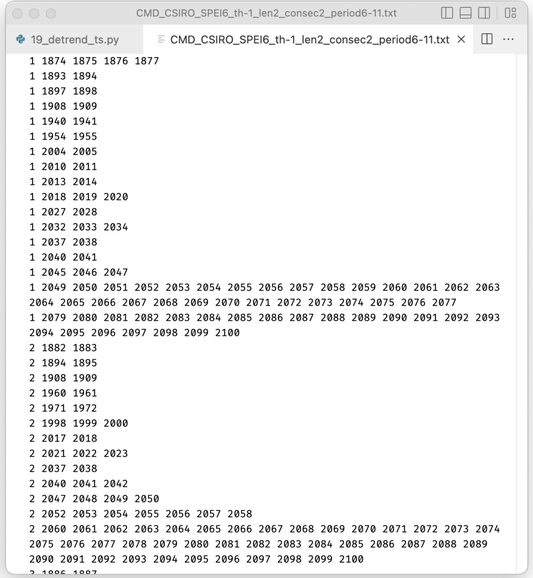
Files (211.5 kB)

| Name | Size | |
|---|---|---|
| 01_regrid_ensemble.py | 2.0 kB | Download |
| md5:362c7d4e63171b1c8fa0ec9556f30ffa | | |
| 02_Makkink_ensemble.py | 1.9 kB | Download |



19_detrend_ts.py

```
import rhine_myd_func as func
import xarray as xr
import numpy as np
import datetime
import matplotlib.pyplot as plt
from numpy.lib import polynomial

poly = 4
type = 'bymonth' #type = 'byyear'

variables = ['pr','tas','rsds']
base_period = [1991,2020]
region_name = 'boxRhine'

for model in func.models:
    for var in variables:

        # open data
        da = func.data_tsensfile(model,var,region_name)
        # detrend
        da_detrend = func.detrend_poly(da,model,base_period,poly,type)
        # save
        da_detrend = da_detrend.astype('float32')
        ds_out = da_detrend.to_dataset()
        if type == 'bymonth':
            file_out = f"{func.datadir}{model}/{var}_m_{model}_{region_name}
            _detrend{poly}M.nc"
        elif type == 'byyear':
            file_out = f"{func.datadir}{model}/{var}_m_ {model}_{region_name}
            _detrend{poly}A.nc"
        ds_out.to_netcdf(file_out)

        # done
        print(f"> {file_out}")
```

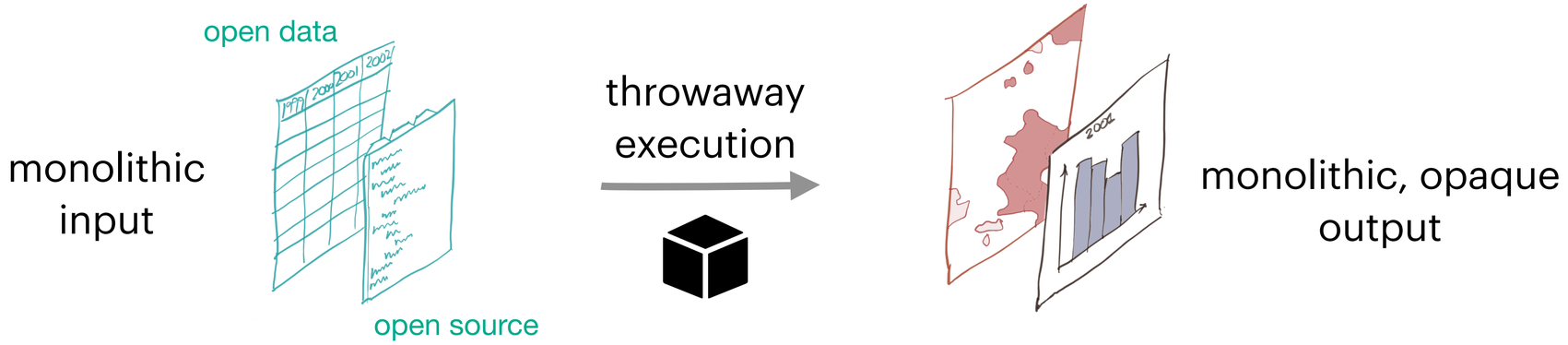


CMD_CSIRO_SPEI6_th-1_len2_consec2_period6-11.txt

```
1 1874 1875 1876 1877
1 1893 1894
1 1897 1898
1 1908 1909
1 1940 1941
1 1954 1955
1 2004 2005
1 2010 2011
1 2013 2014
1 2018 2019 2020
1 2027 2028
1 2032 2033 2034
1 2037 2038
1 2040 2041
1 2045 2046 2047
1 2049 2050 2051 2052 2053 2054 2055 2056 2057 2058 2059 2060 2061 2062 2063
2064 2065 2066 2067 2068 2069 2070 2071 2072 2073 2074 2075 2076 2077
1 2079 2080 2081 2082 2083 2084 2085 2086 2087 2088 2089 2090 2091 2092 2093
2094 2095 2096 2097 2098 2099 2100
2 1882 1883
2 1894 1895
2 1908 1909
2 1960 1961
2 1971 1972
2 1998 1999 2000
2 2017 2018
2 2021 2022 2023
2 2037 2038
2 2040 2041 2042
2 2047 2048 2049 2050
2 2052 2053 2054 2055 2056 2057 2058
2 2060 2061 2062 2063 2064 2065 2066 2067 2068 2069 2070 2071 2072 2073 2074
2075 2076 2077 2078 2079 2080 2081 2082 2083 2084 2085 2086 2087 2088 2089
2090 2091 2092 2093 2094 2095 2096 2097 2098 2099 2100
```

# From open source and open data to “open computation”

We can and should continue to improve open science practices, but there is a problem:




Crucial **relationships** between data and outputs established during execution are lost and are very hard to reconstruct after the fact

My research: can techniques from **programming languages** help us rethink this picture?

**Functional Programs That Explain Their Work**
Perera, Acar, Cheney, Levy **ICFP 2012**

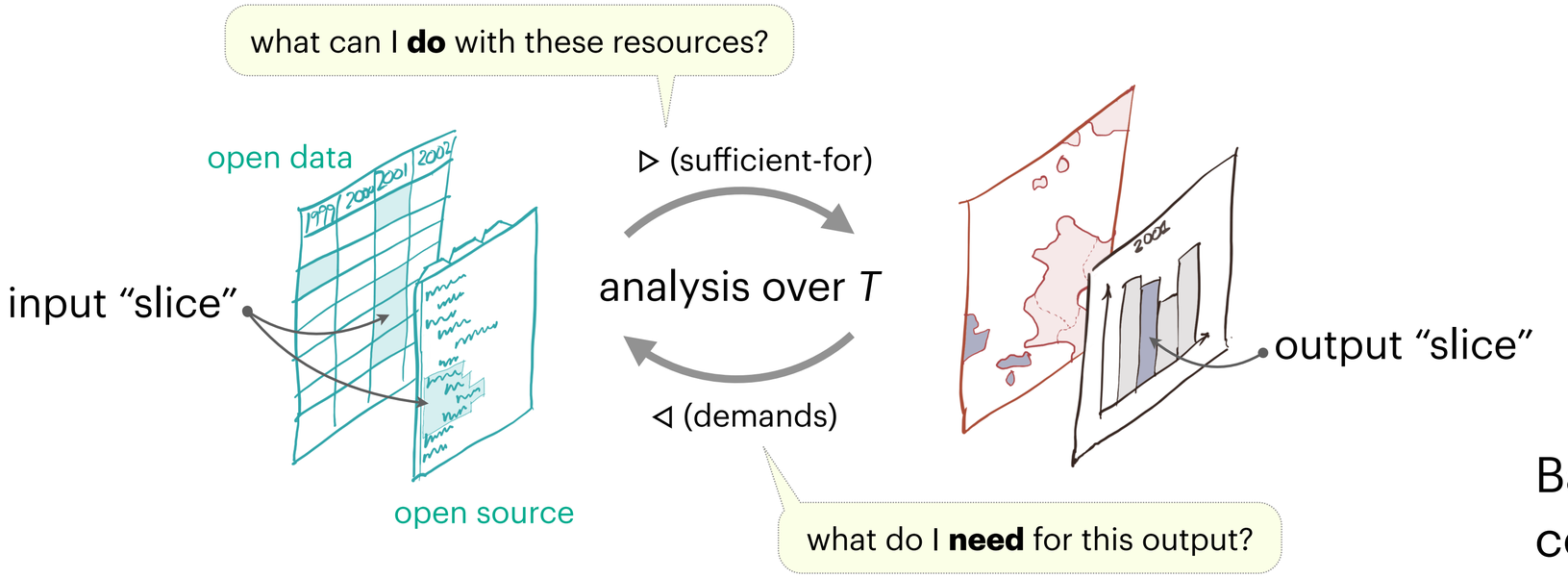
**A Core Calculus for Provenance**
Acar, Ahmed, Cheney, Perera **POST 2012**

# Galois slicing (i)

Replace run-once execution by **persistent and fine-grained bidirectional relationship**

program produces a **trace** *T*
as well as an output

pair of functions $\triangleleft \dashv \triangleright$ relating input and output "slices" called a **Galois connection**




Bake this metadata into computed artefact as a kind of **built-in transparency**

# Demo: “data-transparent” visualisations

Research prototype called **Fluid**

**Causally Consistent Dynamic Slicing**
Perera, Garg, Cheney **CONCUR 2016**

**Imperative Functional Programs That Explain Their Work**
Ricciotti, Stolarek, Perera, Cheney **ICFP 2017**

**Proof-Relevant π-calculus**
Perera, Cheney **MSCS 2018**

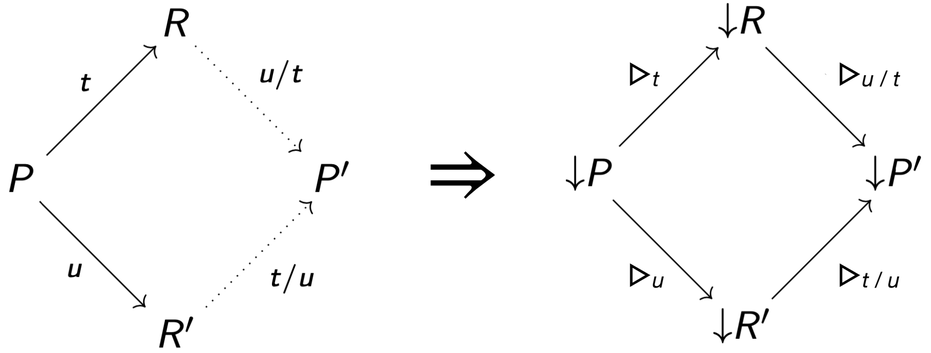
# Galois slicing (ii)

Later work extended to mutable state, exceptions, and **concurrency**

To define ◁ and ▷ for a **specific execution,** need to distinguish:

- **observable** non-determinism (e.g. choice of who to communicate with)
- **unobservable** non-determinism (interleaving of independent steps)

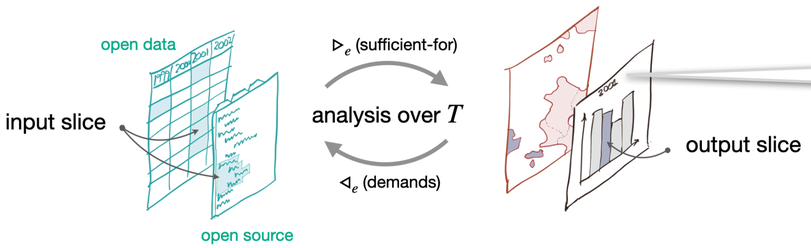
Executions are **causally equivalent** iff they differ only in unobservable choices




Causally equivalent executions induce **equivalent slicing operators** ◁ and ▷

Allows slicing to exploit concurrency whilst producing consistent results

# Linked outputs (i)

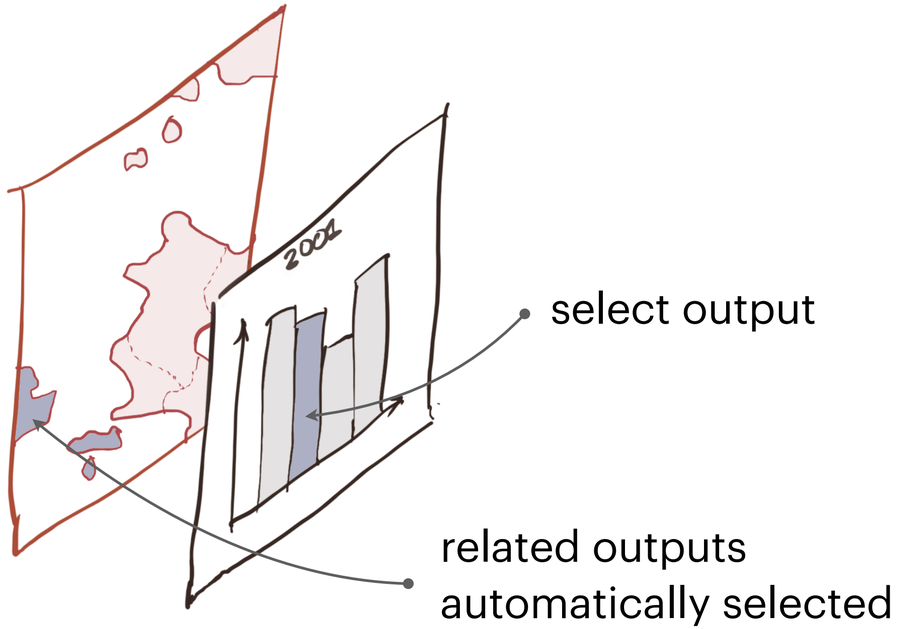



What else might support intelligibility?
What about how **outputs relate to other outputs**?

**Our insight:**
Outputs *x* and *y* are related if they **compete for inputs** i.e. $\triangleleft x$ and $\triangleleft y$ overlap

however: not easy to compute related outputs using only $\triangleleft$ !

**brushing and linking**



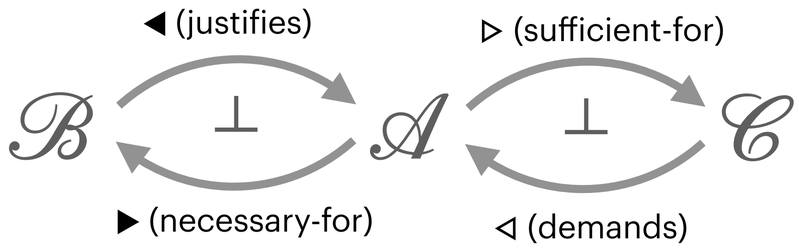

In data visualisation, **brushing and linking** is an important comprehension tool, but ad hoc and poorly understood

# Linked outputs (ii)

Showed how to define **related-outputs** operator ◁▶
that picks out "competing outputs" by allowing slices to have **complements**

**Previous notion of slice**
- value with "holes"
- no notion of complement
- lattice

**New notion ("selection")**
- subset of paths in a value
- complements
- Boolean lattice




**Complements** mean ◁ and ▷
have De Morgan duals

▶ := ¬ ▷ ¬ (**necessary-for**)

◀ := ¬ ▷ ¬ (**justifies**)

# Demo 2: Linked visualisations

# Graphical Galois slicing

with Joe Bond, Minh Nguyen, Cristina David

**Dynamic dependence graphs** (DDGs)
Used for program slicing (Field and Tip 1990)

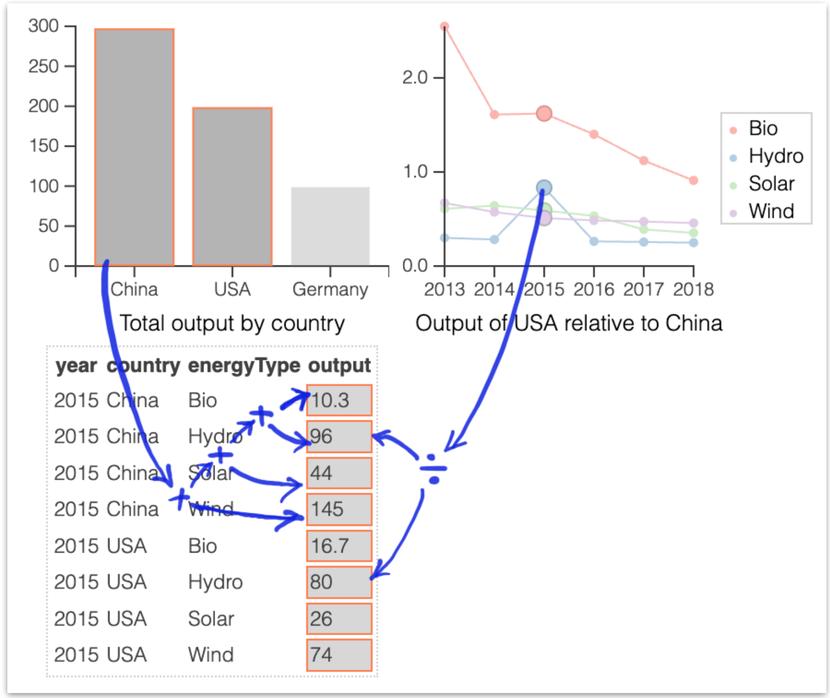
① **Persistent execution** provided by dependence graph $G$ instead of trace

② **Fine-grained analyses** now defined over $G$ (language-agnostic)

▷ sufficient-for

$$\mathscr{P}(V_G) \quad \bot \quad \mathscr{P}(V_G)$$

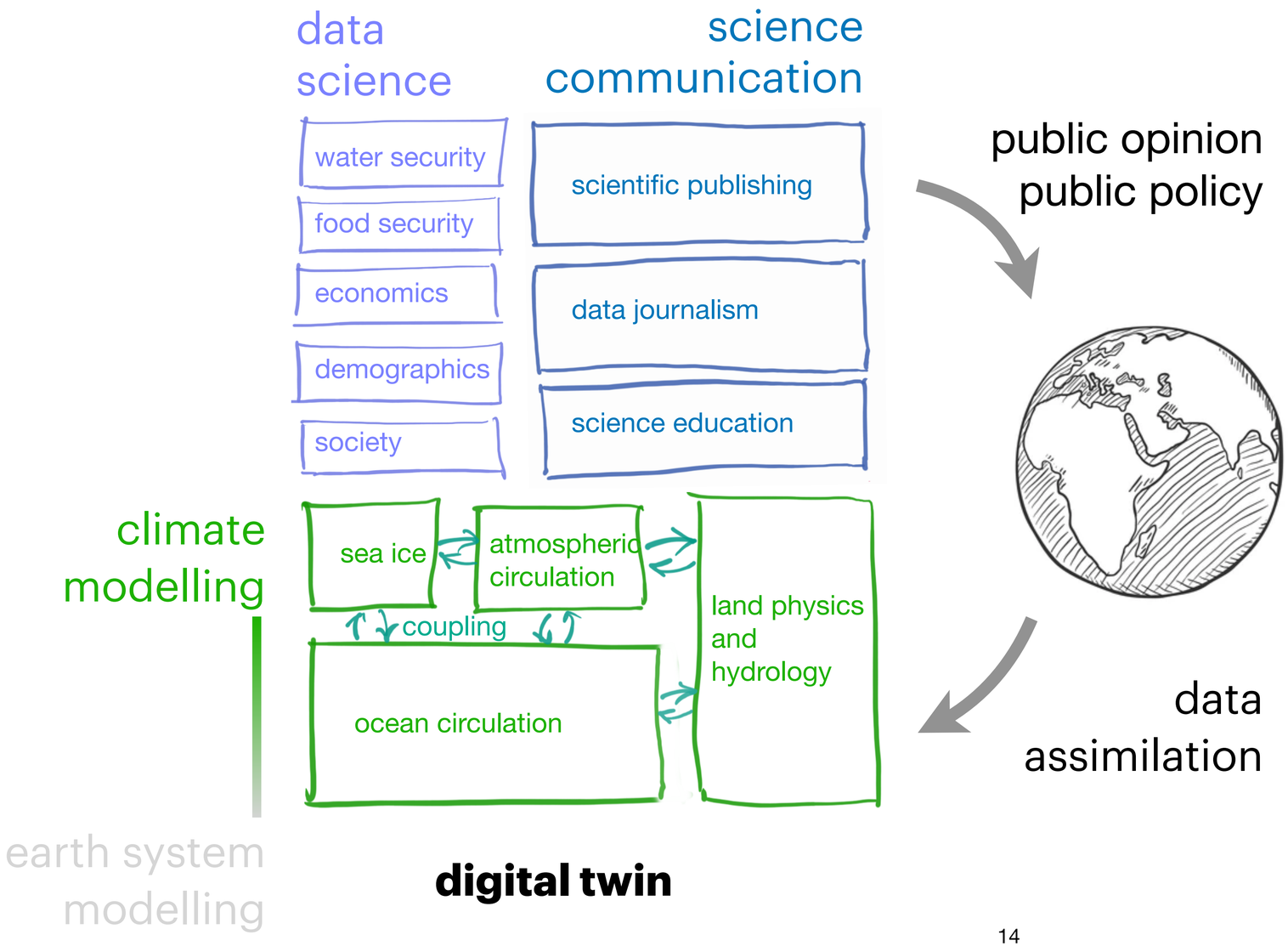
◁ demands (≈ reachability set)




- $G$ directly represents **causal equivalence class** so no rewinding of irrelevant steps
- asymptotically faster in common parallel cases
- **stages** execute concurrently

# Towards transparent computing for climate science

# Computing for climate science




**Transparent software** would improve

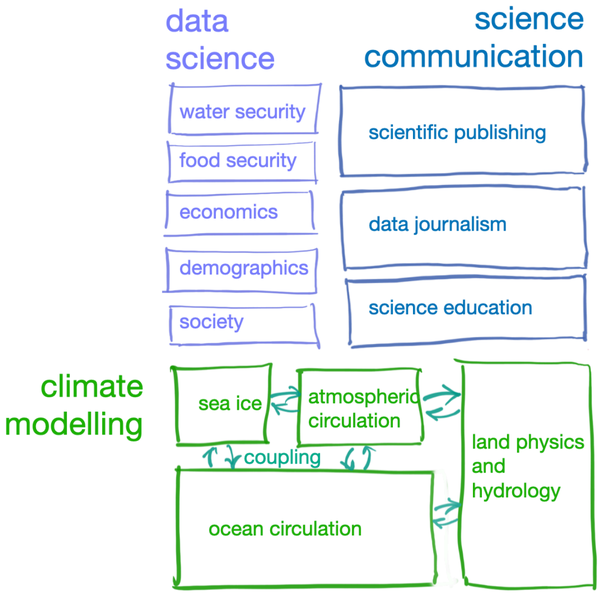
- peer review
- reproducibility
- reuse
- diagnostics

More **efficient** science

- inclusivity
- participation
- accountability
- trustworthiness

More **responsible** science

# However: some challenges




Value proposition of pervasive transparency is clear

But hard to achieve in **heterogeneous, computationally demanding** systems

**multi-paradigm**

numerical simulation, probabilistic, ML, data analysis, streaming data, big data, visualisation, ...

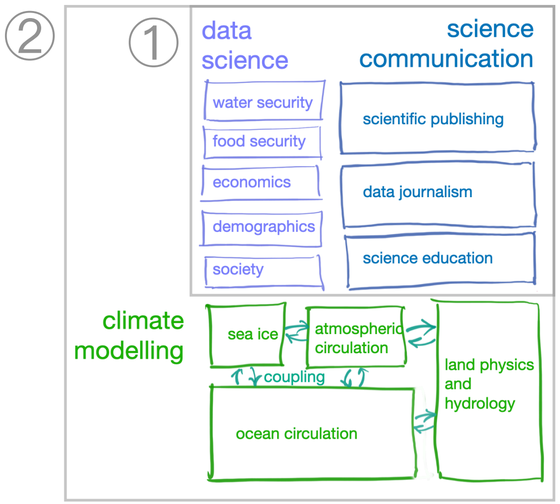
**multi-language**

C/C++, Fortran, R, Python, Julia, JavaScript, SQL, ...

# So, what to do? Two research strategies

① Focus efforts where such features are most valuable and least costly: **climate science communication**

- transparency at a premium
- computationally simpler
- less established tools




② More generally: design explicitly for **gradual adoption** and **scalability**

- techniques for interoperable, partial & approximate transparency in heterogeneous systems

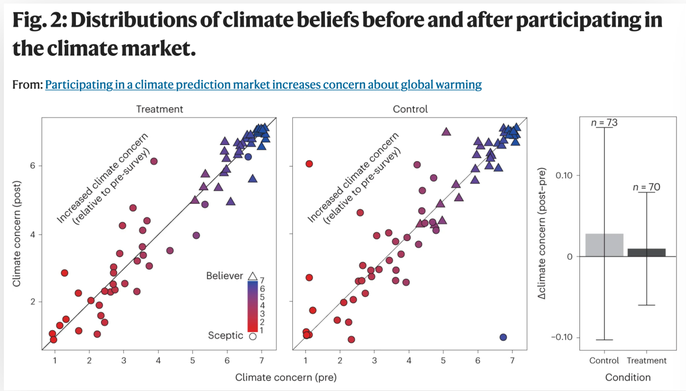
# Transparent research artefacts (i)

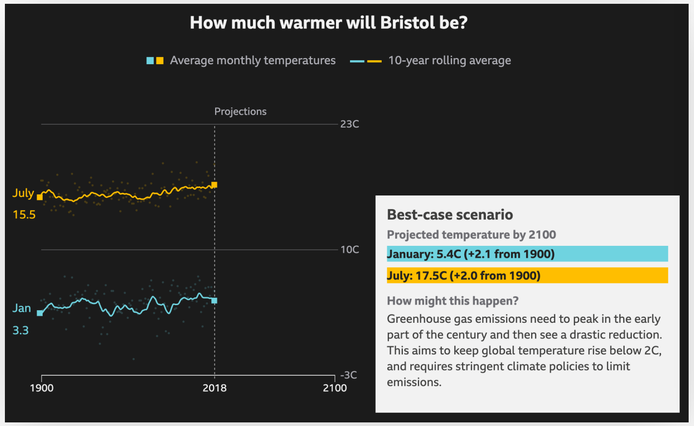
**Today:** papers, news articles and educational tools are opaque

**Goal:** online artefacts which are **transparent and self-explaining** via rich automated interactions, building on Fluid prototype

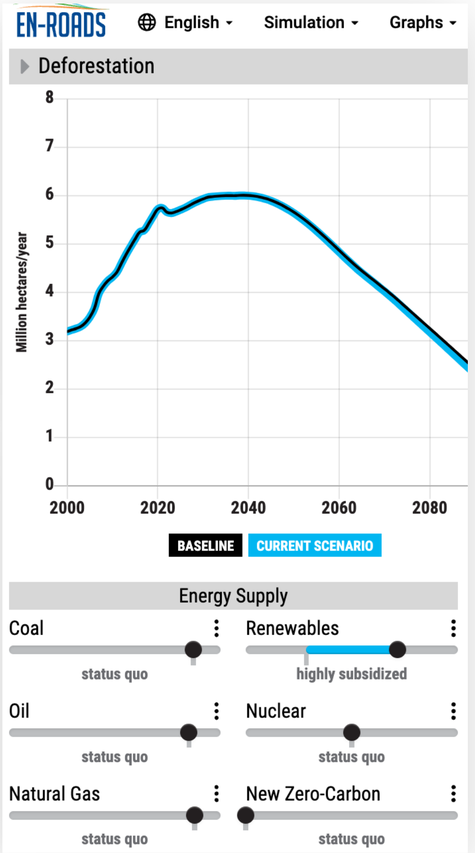
## Climate science communication today

research papers



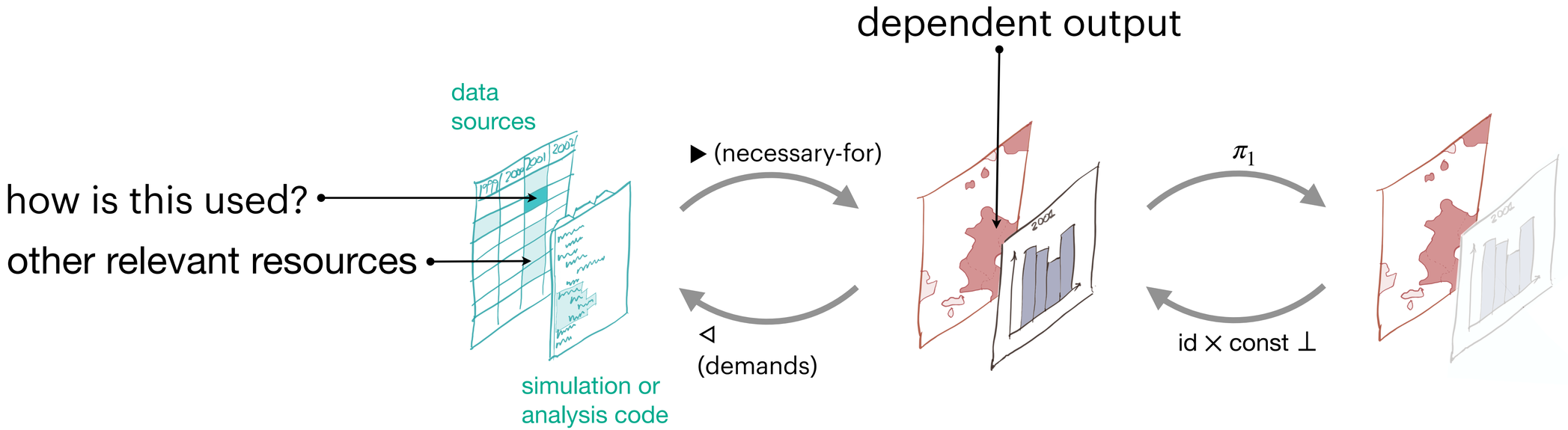

visual journalism




interactive simulation

# Transparent research artefacts (ii)

Automated interactions to support **comprehension** and **reuse**




**related-inputs** := ▶◁ (demands ∘ necessary-for)

Mostly engineering/UX work

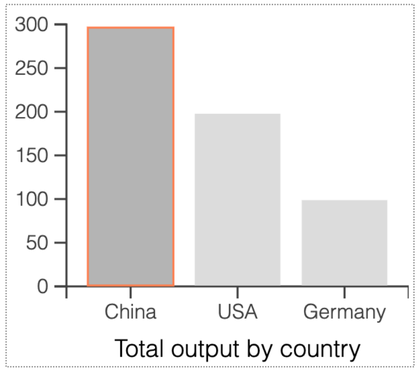
Other adjoint operators for focusing views on **geographical regions** or **time slices**

# Transparent research artefacts (iii)

**Self-explaining computation:** equip output selections with intensional explanations (how, not just what)

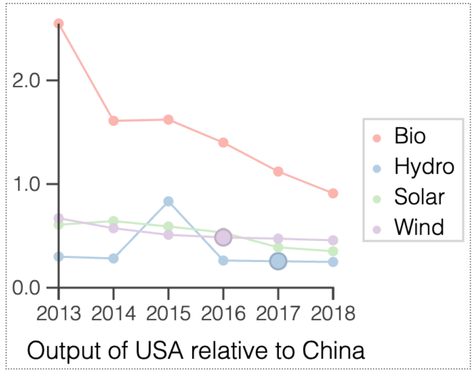
Related work:
executable slicing, expression provenance

Some challenges
- naming and reference
- domain-specific explanations




| *energyType* | *output* |
|---|---|
| Bio | 10.3 |
| Hydro | 96 |
| Solar | 44 |
| Wind | 14.5 |
| **Sum** | **164.8** |

- rollup along dimension

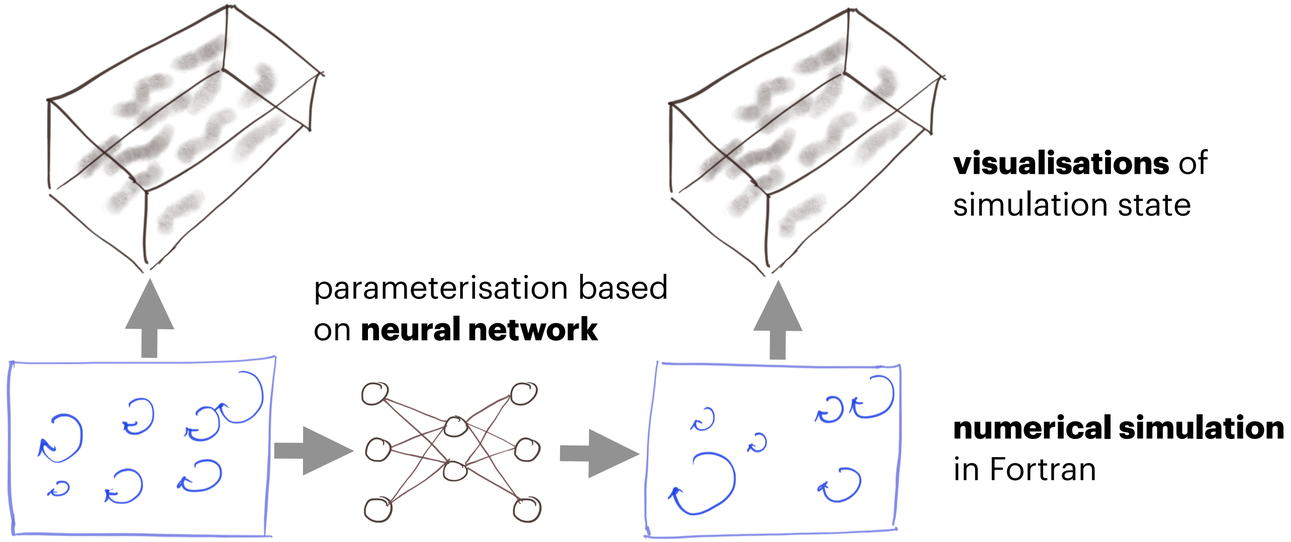



| *year* | *energyType* | *China* | *USA* | *=USA/China* |
|---|---|---|---|---|
| 2016 | Wind | 169 | 82 | 0.49 |
| 2017 | Hydro | 313 | 80 | 0.26 |

- pivoting
- column-wise operations

Use to implement climate articles and explorable simulations like En-ROADS that are **transparent and self-explanatory**

# Gradual, interoperable transparency (i)




**Example**

Subgrid-scale “parameterisations”: pluggable components of climate simulations

How to integrate XAI for neural network with other explainable components?

**Back to research direction** ②

- complex, heterogeneous systems
- transparency solutions need to **interoperate** and **scale**

**Research approach**

- treat gradual, interoperable explainability as first-class problem

# Gradual, interoperable transparency (ii)

**ExPLan:** Explainable Programming Languages
Project lead: James Cheney
Submitted to ERC

## Partial and interoperable transparency

### Interoperable

- Compositionality of explainability notions (e.g. quantitative with symbolic)
- Round-tripping properties
- Language interoperability (FFIs)

### Partial and gradual

- Flexible instrumentation: disable in production code, retain for debugging or teaching
- Approximate (trade precision for performance)
- Partial (trade intensional information)

## Partial transparency

Given dependence graph $G = (V, E)$
if we elide internal nodes but preserve connectivity, resulting graph will yield the same Galois connection $\triangleleft \dashv \triangleright : \mathscr{P}(V) \to \mathscr{P}(V)$

Points to a way to give up intensional explainability for performance

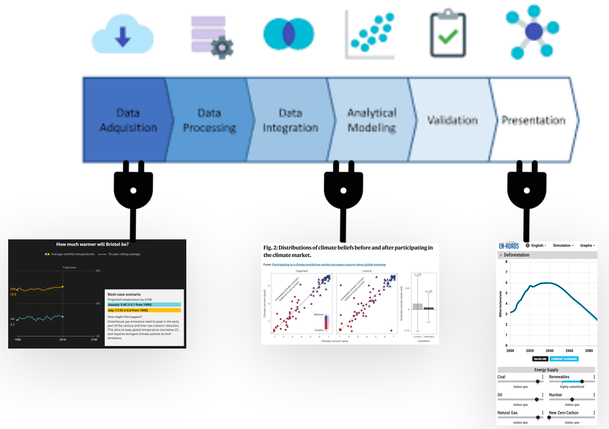
# Conclusion: open climate science needs “open computation”

**Accessible**, **intelligible**, **assessable** and **(re)usable** computational science needs better practices but also **programming languages that are more transparent**

Climate science presents specific challenges to this idea (but also the perfect test bed)

More than anywhere else we want:

- engaged citizens
- transparent governance
- accountable policy makers
- intelligible science




With developments in PL infrastructure we can gradually replace **opaque snapshots** with **transparent views of actual workflows** (different views for different stakeholders)

# Thank you 🥰

Roly Perera
Institute of Computing for Climate Science, University of Cambridge
School of Computer Science, University of Bristol

Institute of
Computing for
Climate Science