Systems in Geospatial Data

WTH is Systems Thinking?

Most geospatial phenomenon is a system

- City's traffic congestion
  - Individual choices of residents
  - Land-use design
  - Public transport availability
- Farmer's crop choice
  - Market incentives, regulation and demand

Why model systems?

- Understand
  - Both overall behavior as well as cause and effect relationships
- Predict

Example

- Flood model
  - How bad floods could be
- Agent based model for behavior of residents
  - What will people decide?
- Migration model
  - Which places might they choose to move to?

Tierolf, Lars, Toon Haer, WJ Wouter Botzen, Jens A. de Bruijn, Marijn J. Ton, Lena Reimann, and Jeroen CJH Aerts. "A coupled agent-based model for France for simulating adaptation and migration decisions under future coastal flood risk." *Scientific Reports* 13, no. 1 (2023): 4176.

What scale to model?

- Spatial scale
  - Local to global
  - Urban sprawl to climate change
- Temporal scale
  - Minute by minute or year by year?

Challenges in modeling scale

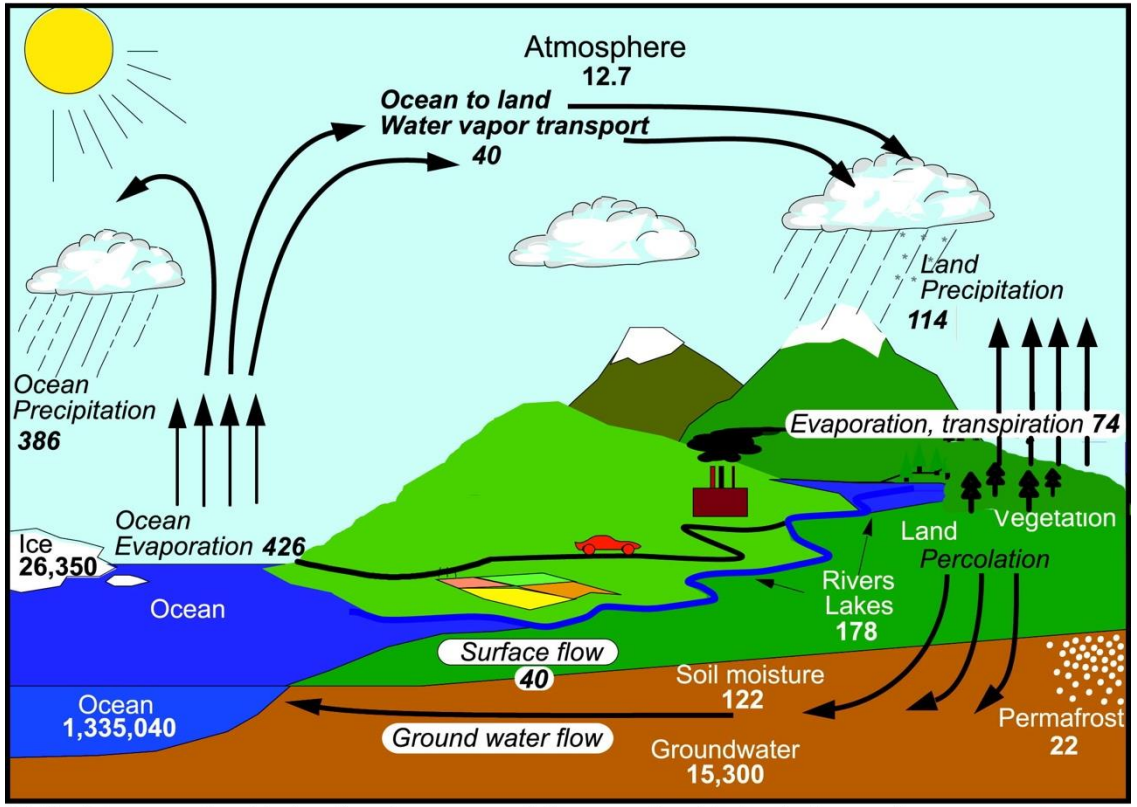
- Links across scales are required
- Nesting of models is required – e.g., climate prediction

Types of models

- Process based (physical/mechanistic models)
- Agent based
- Empirical models
- Statistical models
- Hybrid models
- Cellular Automata/GIS based

Process based models

- Modeling underlying geophysical/biological phenomenon mathematically
- Example
  - Hydrological models - rainfall infliltration, surface runoff and river flow calculations, e.g., SWAT or HEC-HMS
  - Climate models – solve fluid dynamics and thermodynamics for ocean and atmospheric circulation studies respectively



Units: Thousand cubic km for storage, and *thousand cubic km/yr* for exchanges *1990s

Source: https://www.metlink.org/resource/the-changing-water-cycle/

Pros and Cons

Pros

- Can model things in great detail through deterministic equations
- Capture cause and effect mechanisms
  - 'What-if' scenarios

Cons

- Lots of parameters required
- Solving equations can be slow (differential equations)
  - Require a lot of compute resources
- Discretization (choosing a grid cell size) is necessary but non-ideal
  - Need to average subgrid samples
  - Calibration required across broader extents
- Expertise required to use tools

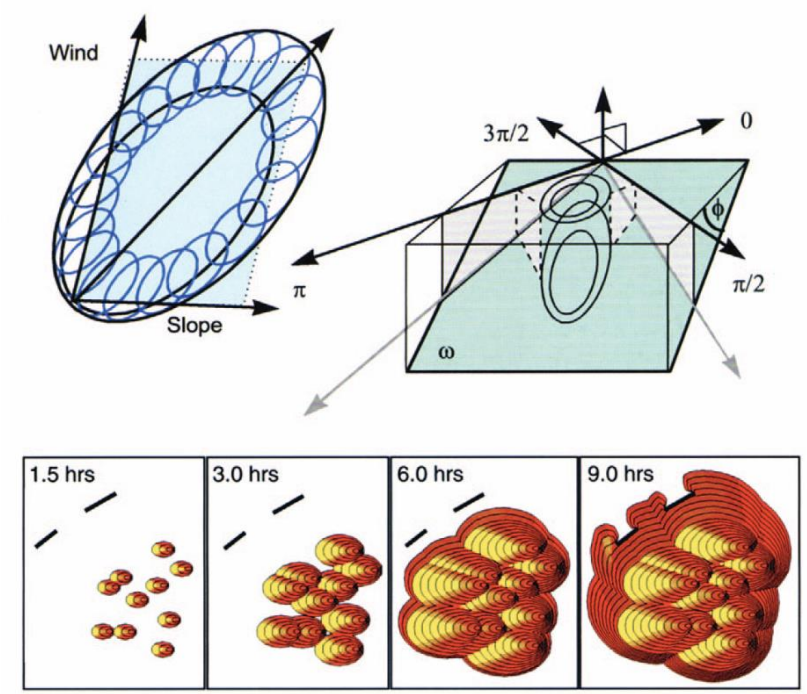
Example use cases

- Tools
  - SWAT – Soil and Water Assessment tools
  - HEC-HMS (Hydrologic engineering center Hydrologic modeling system)
- Studies
  - Assess impact of land use changes on a river basin
  - Predict global water availability due to climate change (General Circulation Models)

Rajib, Adnan, and Venkatesh Merwade. "Hydrologic response to future land use change in the Upper Mississippi River Basin by the end of 21st century." *Hydrological Processes* 31, no. 21 (2017): 3645-3661.

Examples – agriculture

- Crop Growth Models (DSSAT) – simulates growth for 42 types of crops
- Wildfire spread models – FARSITE – utilize physics of combustion and heat transfer to simulate fire behavior across landscapes



Source: https://www.fs.usda.gov/rm/pubs/rmrs_rp004.pdf

https://www.fao.org/land-water/land/land-governance/land-resources-planning-toolbox/category/details/en/c/1027490/

Examples – health (Epidemiology)

- SIR (Susceptible – Infected – Recovered) disease model
- Extended with spatial diffusion
  - Spatiotemporal process model of disease spread

Agent-based models

Agent-based modeling (ABM)

- A bottom-up modeling approach
  - Simulate actions and interactions of individual agents
  - Macro-scale patterns emerge out of individual behavior
  - Cars moving through city or individuals making habitation choices
- Environment
  - Grid (raster), Network (graph) or Continuous coordinates (equations)

Urban Planning

- Modeling urban sprawl
  - Model developers and households as agents making location choices
    - SIMPOP model (https://journals.sagepub.com/doi/10.1068/b240287)
    - UrbanSim (https://github.com/UDST/urbanism)
  - Human migration under climate change
    - Households in Bangladesh deciding to relocate inland given flooding risk

Paul, Bimal Kanti, Munshi Khaledur Rahman, Max Lu, and Thomas W. Crawford. "Household migration and intentions for future migration in the climate change vulnerable lower Meghna estuary of coastal Bangladesh." *Sustainability* 14, no. 8 (2022): 4686.

Disease outbreaks

- COVID-19 outbreaks
  - Test lockdown policies by simulating how contact patterns affect virus transmissions
- Agent-based cholera model in Kumasi, Ghana

Augustijn, Ellen-Wien, Tom Doldersum, Juliana Useya, and Denie Augustijn. "Agent-based modelling of cholera diffusion." *Stochastic environmental research and risk assessment* 30 (2016): 2079-2095.

Roy, Shovonlal. "COVID-19 pandemic: Impact of lockdown, contact and non-contact transmissions on infection dynamics." *MedRxiv* (2020): 2020-04.

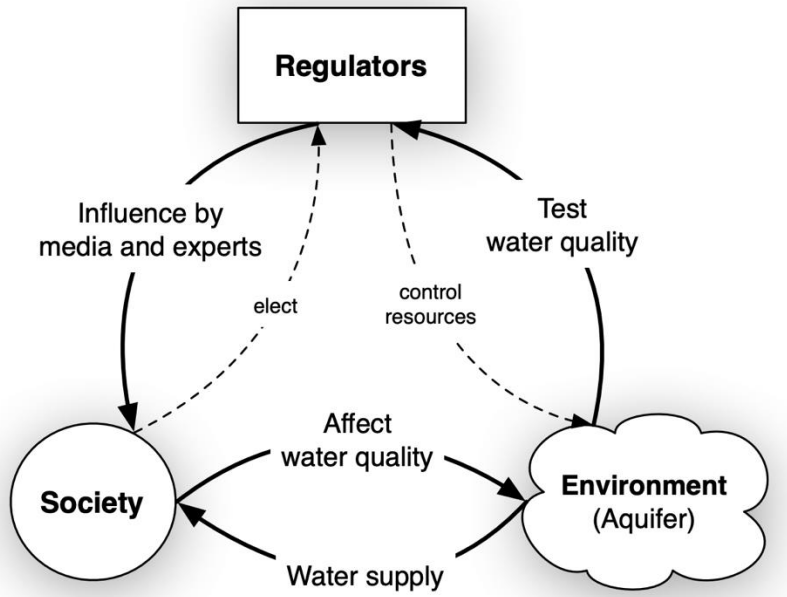
Agriculture

- Geospatial pest management ABM
  - Forest insect infestation (emerald ash borer (EAB)) and its biocontrol agent

Spatial Scales of studies

- Typical usage
  - Regional scale simulations – urban disease/traffic
- Large scale ABMs
  - Global trade models with agents as countries
  - Macro-scale epidemic models
  - Socio-hydrological models

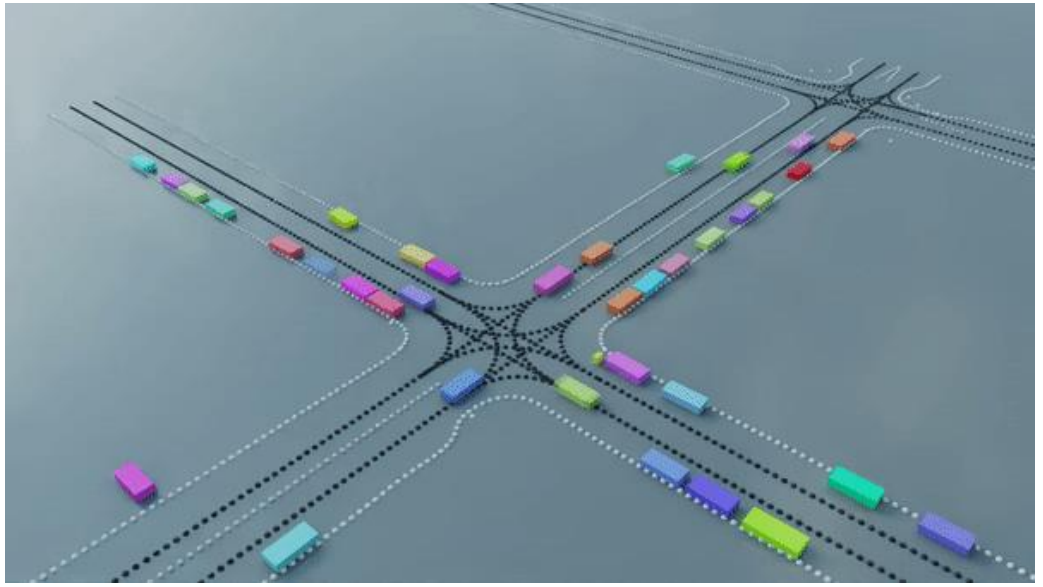
GIS data is used to ground simulations (realistic environments)



Bakarji, Joseph, Daniel O'Malley, and Velimir V. Vesselinov. "Agent-based socio-hydrological hybrid modeling for water resource management." *Water Resources Management* 31 (2017): 3881-3898.

Strengths

- Emergent phenomenon with simple rules
- No need for ideal behavior assumptions
  - Supports heterogeneity
  - Flexible modeling (including integration of spatial context)
- Great visuals



Limitations

- Agent rules are not trivial
- Complex rules need more compute to simulate
- Risk ‘ad-hoc’ tweaking of agent rules to calibrate results
- Determinism across multiple iterations, specially, for stochastic rules is not guaranteed

Empirical models

Learning from past data

- Function or decision rule inferred directly from data
- Empirical modeling of the distribution from which the data is generated
  - Geostatistical Interpolation (Kriging)
    - LULC, Wildfire risk, Flood Susceptibility, etc.
- Physics-inspired ML (mix of process models and empirical models)

Müller, Jörg, Oliver Mitesser, H. Martin Schaefer, Sebastian Seibold, Annika Busse, Peter Kriegel, Dominik Rabl et al. "Soundscapes and deep learning enable tracking biodiversity recovery in tropical forests." *Nature communications* 14, no. 1 (2023): 6191.

Spatial scales of studies

- Regional models (due to local data collection)
- Global models (for satellite/climatic variable based data)
- Geospatial generalization is a big concern
  - Geospatial cross-validation must be followed while verifying results
  - Don't sample test and train from similar locations

Pros and Cons

Pros

- Flexible and powerful models available
- Faster inferences as compared to process models
- Improve with more data
- Easy to operationalize

Cons

- Dependence on data quality and scope
- Lack of causality
- Overfitting is common

Statistical models

Why statistical models?

- Predict the underlying process parameters
  - Ability to provide confidence intervals, hypothesis tests
  - Examples: Linear Regression, Generalized Linear Models (GLMs), **Geostatistical models (kriging, variograms)**, Spatial autoregressive models
  - Linear relationship assumption + epsilon (probabilistic form)
  - **Point process models**
    - Earthquakes, crimes, etc.
  - **Spatial interaction models**
    - Gravity models that determine migration or trade flow between areas

$$F_{ij} = C\frac{GDP_i \bullet GDP_j}{D_{ij}},$$

Example of gravity model used in trade between two countries

Kriging

Comes from gold samples in African mines

Step 1:
Create a random process that best justifies the data x1 to xn

Step 2:
For locations x1 to xn, get values z(x1) to z(xn), also called realizations of these random variables

Assumption 1: First moment stationarity
Value = mean(all values)

Assumption 2: Second moment stationarity
Value depends only on distance from a certain sample

The hypothesis of stationarity related to the second moment is defined in the following way: the correlation between two random variables solely depends on the spatial distance between them and is independent of their location. Thus if $\mathbf{h} = x_2 - x_1$ and $h = |\mathbf{h}|$, then:

$$C\big(Z(x_1), Z(x_2)\big) = C\big(Z(x_i), Z(x_i + \mathbf{h})\big) = C(h),$$
$$\gamma\big(Z(x_1), Z(x_2)\big) = \gamma\big(Z(x_i), Z(x_i + \mathbf{h})\big) = \gamma(h).$$

For simplicity, we define $C(x_i, x_j) = C\big(Z(x_i), Z(x_j)\big)$ and $\gamma(x_i, x_j) = \gamma\big(Z(x_i), Z(x_j)\big)$.

This hypothesis allows one to infer those two measures – the variogram and the covariogram:

$$\gamma(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} \big(Z(x_i) - Z(x_j)\big)^2,$$

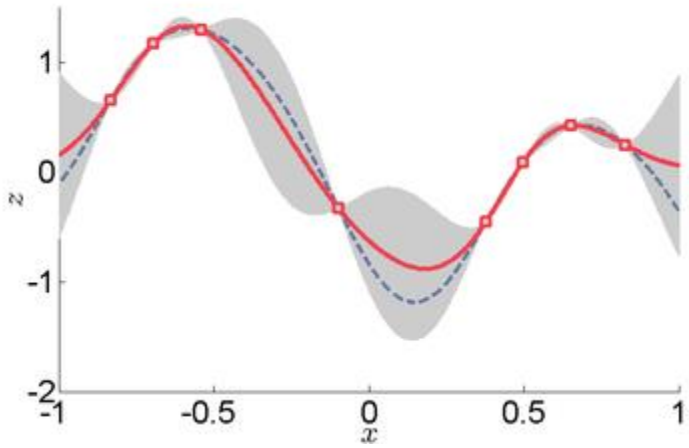
$$C(h) = \frac{1}{|N(h)|} \sum_{(i,j)\in N(h)} \big(Z(x_i) - m(h)\big)\big(Z(x_j) - m(h)\big),$$

where:

$$m(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} Z(x_i) + Z(x_j);$$

$N(h)$ denotes the set of pairs of observations $i,\ j$ such that $|x_i - x_j| = h$, and $|N(h)|$ is the number of pairs in the set.

In this set, $(i,\ j)$ and $(j,\ i)$ denote the same element. Generally an "approximate distance" h is used, implemented using a certain tolerance.



Underlying a prior which is adjusted to the samples derived in the distribution

Examples

- Building house price prediction (spatial regression)
  - Incorporate spatial autocorrelation
- Rainfall or pollution maps
  - Kriging apart from the sparse sampling
- Hydrology (ground water)
- Agriculture (soil nutrient)
- Disease mapping

Spatial scale

- Where was the data collected from determines the scale
- Kriging/Regression is typically regional

Pros

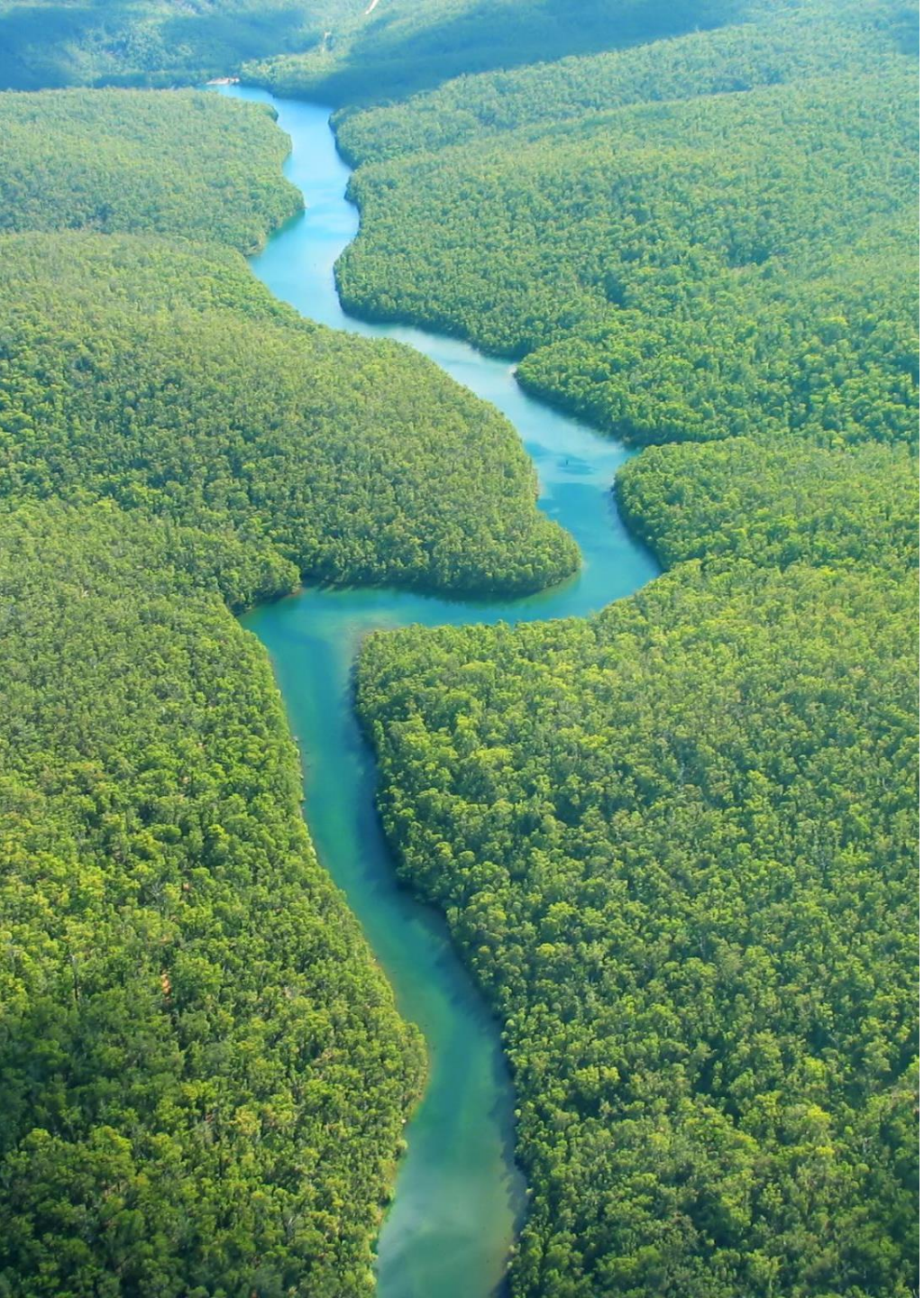
- Rigorous uncertainty qualification
- Spatial dependency incorporation
- Mature methodology
- Explainable methods

Cons

- Rigid functional forms
- Often come with assumptions on normality or stationarity
- Sensitive to scale
- Limited dynamics (often target modeling of steady state)

Hybrid Models

Also known as multi-paradigm and integrated modeling



Why hybridize?

- Geospatial systems involve both rivers and incomes, forests and castes.
- Some of it you can model with physics, other is socio-political.
- Therefore, you require different approaches to come together, e.g., process-based environment model with an agent based socio-economic model

Some examples

- Agent based systems determining farmer's choices + process based models for hydrological outcomes (RiverWare.org)
  - Model human risk perception and decisions along with its outcomes on rivers, dams, etc.
- Land-use allocation models
  - Macro scale land-use allocation (economics based) + Small scale spatially aware land-use allocation
  - CLUE – Conversion of Land-use and its effect
  - SLEUTH - **Slope, Land use, Exclusion, Urban extent, Transportation and Hillshade**

Hua, Lizhong, Lina Tang, Shenghui Cui, and Kai Yin. "Simulating urban growth using the SLEUTH model in a coastal peri-urban district in China." *Sustainability* 6, no. 6 (2014): 3899-3914.

Ewers, M. "Combining hydrology and economics in a system dynamics approach: modeling water resources for the San Juan Basin." In *Proc., 23rd International Conference of the System Dynamics Society, July*, vol. 1721. 2005.

Examples

- A disease progression model within an agent following SEIR differential equations + interactions set up through an agentic model
- Internal viral load dynamics - Process + contact networks - ABM (COVID-19)
- Physics informed ML that simulates particle dispersion with physics based loss functions
- Crop yield forecasting that combines crop growth model with parameters being predicted with an ML model for various different climate scenarios

Spatial scales

- Global climate model + local economic model
- Parcel level land decisions + regional market prices
- Building energy process models + agent based behavioral models -> city scale models
- Power grid simulations (EPA: https://www.epa.gov/power-sector-modeling/post-ira-2022-reference-case)

Pros

- Comprehensive
- Complementary strengths
- Cross-scale linkages
- Lots of customization and innovation possible

Cons

- Calibration and validation is non-trivial
- Transparency and stakeholder understanding is lower
- Maintenance and reproducibility is complex

Let’s study a system

- Use your favorite physics!