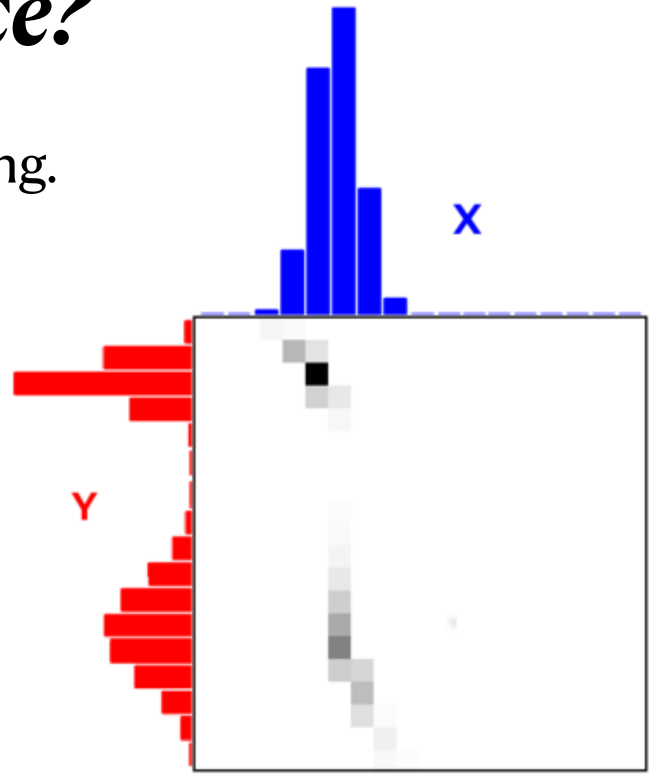
What is… The Wasserstein Distance?

An introduction, with application to climate modelling.

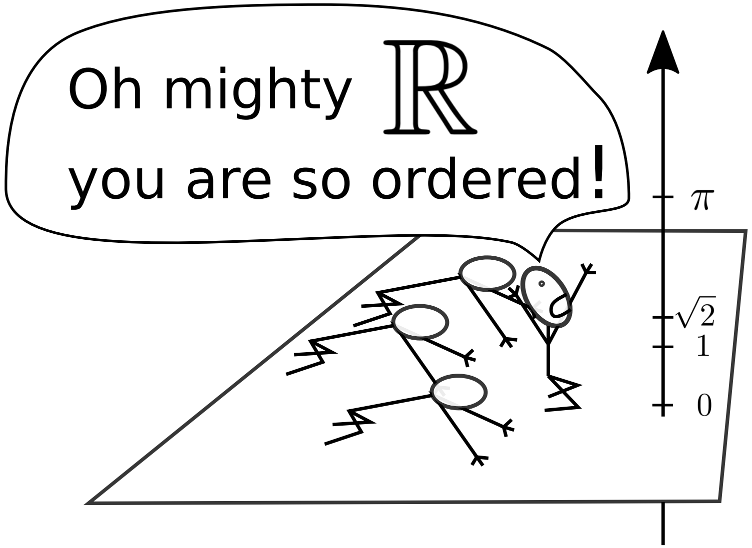
(joint with Mat Chantry, Milan Klöwer & Tim Palmer)

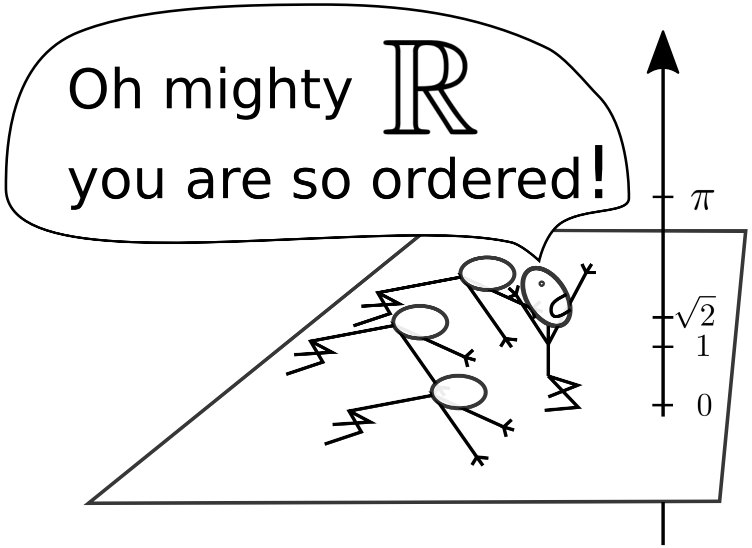


E. Adam Paxton

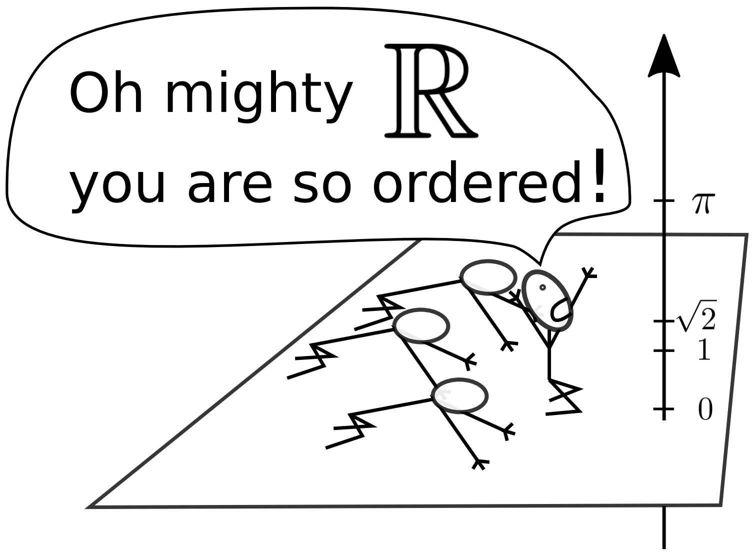
Predictability group internal seminar 09.11.20

UNIVERSITY OF OXFORD

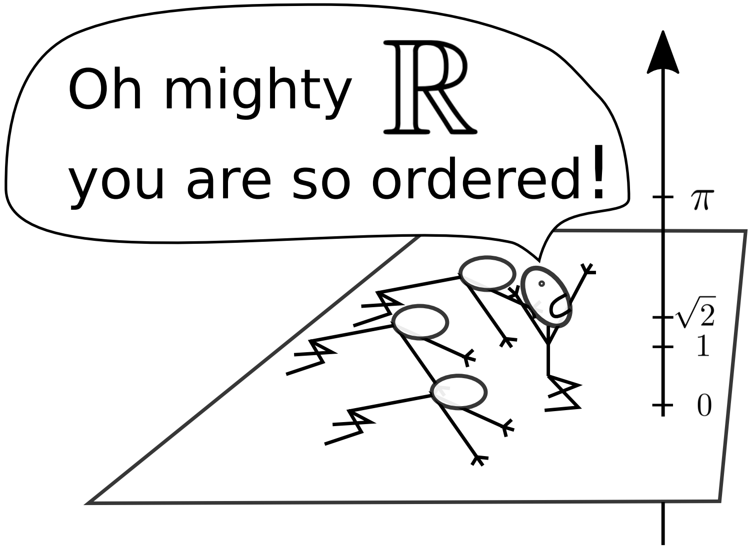

Oh mighty $\mathbb{R}$
you are so ordered!
π
$\sqrt{2}$
1
0


Oh mighty $\mathbb{R}$
you are so ordered!
π
$\sqrt{2}$
1
0


My country is the best! We have the *highest GDP.*
Actually, mine is the best. We have the *longest ski-slope.*
I think my country is best. We have the *most biodiversity.*
???...



- **Real world problems are multi-dimensional.**



11° Kilgetty, GB >

Fri, Oct 09, 2020

Newsweek

U.S. | World | Business | Tech & Science | Culture | Newsgeek | Sports | Health | The Debate

WORLD

How China Buried the Green GDP

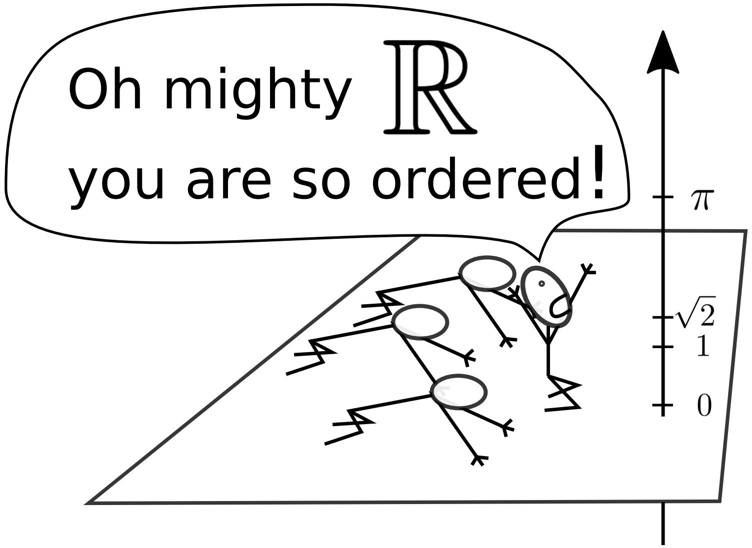
BY MELINDA LIU ON 6/28/08 AT 8:03 AM EDT

SHARE

WORLD

Ask Chinese officials why their nation's environment is so toxic; you'll get a list of scientific-sounding explanations. The population is huge and dense. Arable land per capita is alarmingly sparse. Despite stunning rates of economic growth, many Chinese remain poor and rural, prone to ungreen behaviors such as tossing pollutants and trash into the rivers. But the real question is why China fares poorly in Yale and

- **Real world problems are multi-dimensional.**



11° Kilgetty, GB >

Fri, Oct 09, 2020

Newsweek

U.S. | World | Business | Tech & Science | Culture | Newsgeek | Sports | Health | The Debate

WORLD

How China Buried the Green GDP

BY MELINDA LIU ON 6/28/08 AT 8:03 AM EDT

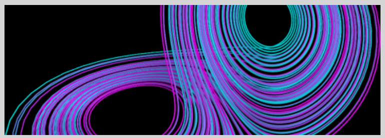
SHARE

WORLD

Ask Chinese officials why their nation's environment is so toxic; you'll get a list of scientific-sounding explanations. The population is huge and dense. Arable land per capita is alarmingly sparse. Despite stunning rates of economic growth, many Chinese remain poor and rural, prone to ungreen behaviors such as tossing pollutants and trash into the rivers. But the real question is why China fares poorly in Yale and

- **Real world problems are multi-dimensional.**
- If anyone says that their metric is the *best*, you should probably be cynical!

I'm now going to tell you that the Wasserstein Metric is the best way to measure distance between probability distributions.



I'm now going to tell you that the Wasserstein Metric is the best way to measure distance between probability distributions.

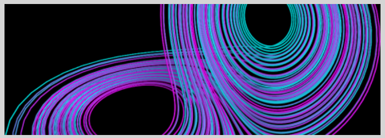
Plan of talk:

1. What is the Wasserstein distance?
2. What are the advantages of the WD, and how to compute it.
3. An application: exploring model climatology in low-precision.

1) What is the Wasserstein Distance?

1) What is the Wasserstein Distance?

- The WD (Earth Mover's distance) is a distance between probability distributions (measures) μ & ν.

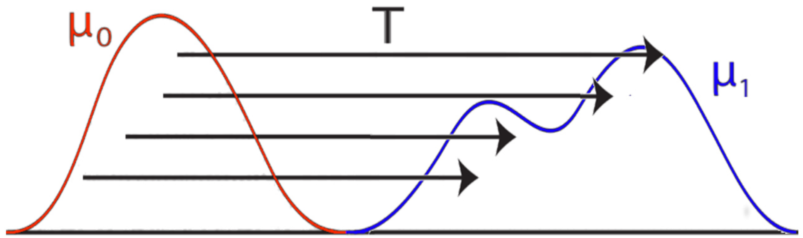


1) What is the Wasserstein Distance?

- The WD (Earth Mover's distance) is a distance between probability distributions (measures) μ & ν.
- It comes from the theory of optimal transport.

1) What is the Wasserstein Distance?

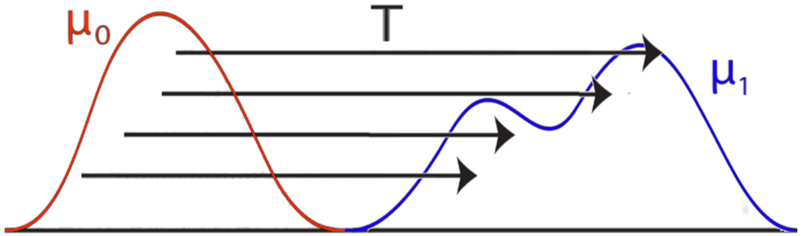
- The WD (Earth Mover's distance) is a distance between probability distributions (measures) μ & ν.
- It comes from the theory of optimal transport.
- Think of μ & ν as *mass distributions*. You are tasked with transporting the mass from μ to ν.



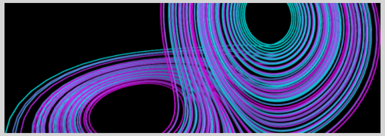
N. Papadakis, Optimal Transport for Image Processing, habilitation à diriger des recherches, Université de Bordeaux, Dec. 2015

1) What is the Wasserstein Distance?

- The WD (Earth Mover's distance) is a distance between probability distributions (measures) μ & ν.
- It comes from the theory of optimal transport.
- Think of μ & ν as *mass distributions*. You are tasked with transporting the mass from μ to ν.
- The cost to transport unit mass from x to y is $c(x, y)$.

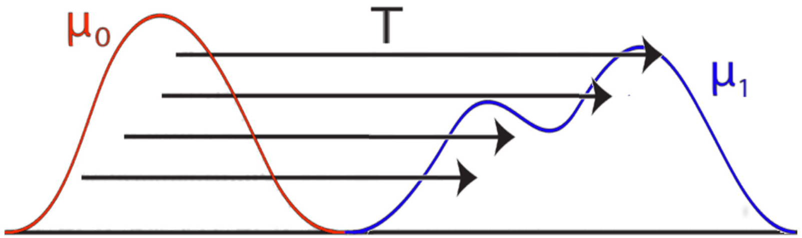


N. Papadakis, Optimal Transport for Image Processing, habilitation à diriger des recherches, Université de Bordeaux, Dec. 2015

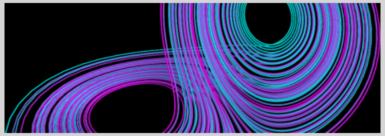


1) What is the Wasserstein Distance?

- The WD (Earth Mover's distance) is a distance between probability distributions (measures) μ & ν.
- It comes from the theory of optimal transport.
- Think of μ & ν as *mass distributions*. You are tasked with transporting the mass from μ to ν.
- The cost to transport unit mass from x to y is $c(x, y)$.
- You want the cheapest strategy.

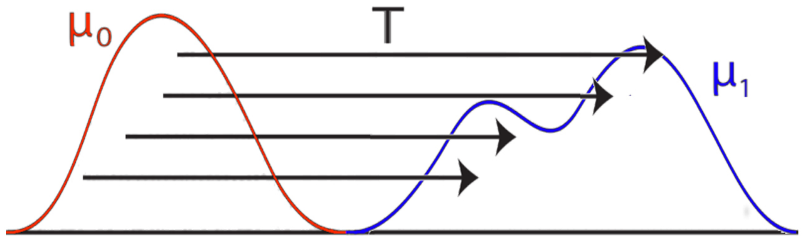


N. Papadakis, Optimal Transport for Image Processing, habilitation à diriger des recherches, Université de Bordeaux, Dec. 2015

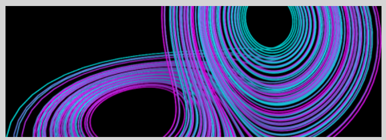


1) What is the Wasserstein Distance?

- The WD (Earth Mover's distance) is a distance between probability distributions (measures) μ & ν.
- It comes from the theory of optimal transport.
- Think of μ & ν as *mass distributions*. You are tasked with transporting the mass from μ to ν.
- The cost to transport unit mass from x to y is $c(x, y)$.
- You want the cheapest strategy.
- For the case $c(x, y) = |x - y|^p$ we call the optimal cost the p-Wasserstein Distance (we'll always take $p = 1$)



N. Papadakis, Optimal Transport for Image Processing, habilitation à diriger des recherches, Université de Bordeaux, Dec. 2015



There are two formulations of Optimal Transport: *Monge* (1781) and *Kantorovich* (1942).

There are two formulations of Optimal Transport: *Monge* (1781) and *Kantorovich* (1942).

Monge's formulation (1781):

- Suppose

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i} \ , \ \nu = \frac{1}{N}\sum_{i=1}^{N} \delta_{y_i} \, .$$

(think discrete, equal masses)

There are two formulations of Optimal Transport: *Monge* (1781) and *Kantorovich* (1942).

Monge's formulation (1781):

- Suppose

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i} \ , \ \nu = \frac{1}{N}\sum_{i=1}^{N} \delta_{y_i} \, .$$

(think discrete, equal masses)

- A *transport strategy* is a permutation of N objects $\sigma \in S_N$.

The cost of a strategy is $\displaystyle\frac{1}{N}\sum_{i=1}^{N} c(x_i, y_{\sigma(i)})$.

$$\boxed{\mathrm{WD}_1(\mu,\nu) := \min_{\sigma\in S_N} \frac{1}{N}\sum_{i=1}^{N} |x_i - y_{\sigma(i)}|}$$

Kantorovich’s formulation (1942):

Kantorovich's formulation (1942):

- Suppose

$$\mu = \sum_{i=1}^{M_1} p_i \delta_{x_i} \,, \qquad \nu = \sum_{j=1}^{M_2} q_j \delta_{y_j}$$

think continuous masses / histograms (can be more general than the above)

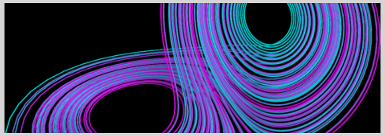
Kantorovich's formulation (1942):

- Suppose

$$\mu = \sum_{i=1}^{M_1} p_i \delta_{x_i}\ , \qquad \nu = \sum_{j=1}^{M_2} q_j \delta_{y_j}$$

think continuous masses / histograms (can be more general than the above)

- A *transport strategy* is a matrix π where π_{ij} is mass transported from i to j



UNIVERSITY OF OXFORD

Kantorovich's formulation (1942):

- Suppose

$$\mu = \sum_{i=1}^{M_1} p_i \delta_{x_i} , \quad \nu = \sum_{j=1}^{M_2} q_j \delta_{y_j}$$

think continuous masses / histograms (can be more general than the above)

- A *transport strategy* is a matrix π where π_{ij} is mass transported from i to j
- By conservation of mass π belongs to $\Pi(\mu, \nu) = \{\pi_{ij} \geq 0 : \sum_j \pi_{ij} = p_i, \sum_i \pi_{ij} = q_j\}$

Kantorovich's formulation (1942):

- Suppose

$$\mu = \sum_{i=1}^{M_1} p_i \delta_{x_i}\ , \qquad \nu = \sum_{j=1}^{M_2} q_j \delta_{y_j}$$

think continuous masses / histograms (can be more general than the above)

- A *transport strategy* is a matrix π where π_{ij} is mass transported from i to j
- By conservation of mass π belongs to $\Pi(\mu, \nu) = \{\pi_{ij} \geq 0 : \sum_j \pi_{ij} = p_i, \sum_i \pi_{ij} = q_j\}$

$$\boxed{\mathrm{WD}_1(\mu, \nu) := \min_{\pi \in \Pi(p,q)} \sum_{i,j} |x_i - y_j| \pi_{ij}}$$

Kantorovich's formulation (1942):

- Suppose

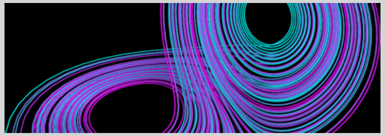
$$\mu = \sum_{i=1}^{M_1} p_i \delta_{x_i} \ , \qquad \nu = \sum_{j=1}^{M_2} q_j \delta_{y_j}$$

think continuous masses / histograms (can be more general than the above)

- A *transport strategy* is a matrix π where π_{ij} is mass transported from i to j
- By conservation of mass π belongs to $\Pi(\mu,\nu) = \{\pi_{ij} \geq 0 : \sum_j \pi_{ij} = p_i, \sum_i \pi_{ij} = q_j\}$

$$\boxed{\mathrm{WD}_1(\mu,\nu) := \min_{\pi \in \Pi(p,q)} \sum_{i,j} |x_i - y_j| \pi_{ij}}$$

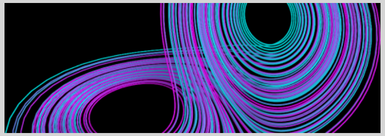
nb. when $M_1 = M_2 = N$ and $p_i = q_i = \frac{1}{N}$ it turns out the two definitions are equivalent.



2) What are the advantages of the WD?

2) What are the advantages of the WD?

(i) It *metrizes* the space of probability distributions.



2) *What are the advantages of the WD?*

(i) It *metrizes* the space of probability distributions.

If μ_k is a sequence of probability distributions, then

$$\mathrm{WD}_1(\mu_k, \mu) \to 0 \quad \textit{if \& only if} \quad \mu_k \to \mu \ \textit{(weak}\star\textit{)}$$

where $\mu_k \to \mu$ (weak$\star$) means:

$$\int_{\mathbb{R}^n} \phi(x) d\mu_k(x) \to \int_{\mathbb{R}^n} \phi(x) d\mu(x) \quad \text{for any bounded function } \phi(x)$$

2) *What are the advantages of the WD?*

(i) It *metrizes* the space of probability distributions.

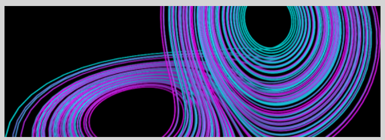
If μ_k is a sequence of probability distributions, then

$$\mathrm{WD}_1(\mu_k, \mu) \to 0 \quad \textit{if \& only if} \quad \mu_k \to \mu \ \textit{(weak}\star\textit{)}$$

where $\mu_k \to \mu$ (weak$\star$) means:

$$\int_{\mathbb{R}^n} \phi(x) d\mu_k(x) \to \int_{\mathbb{R}^n} \phi(x) d\mu(x) \quad \text{for any bounded function } \phi(x)$$

nb. If you don't know this notation, think $d\mu(x) = f(x)dx$ where f is a PDF.



2) What are the advantages of the WD?

(i) It *metrizes* the space of probability distributions.

If μ_k is a sequence of probability distributions, then

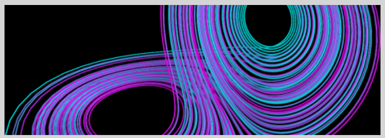
$$\mathrm{WD}_1(\mu_k, \mu) \to 0 \quad \textit{if \& only if} \quad \mu_k \to \mu \ (\textit{weak}\star)$$

where $\mu_k \to \mu$ (weak$\star$) means:

$$\int_{\mathbb{R}^n} \phi(x) d\mu_k(x) \to \int_{\mathbb{R}^n} \phi(x) d\mu(x) \quad \text{for any bounded function } \phi(x)$$

nb. If you don't know this notation, think $d\mu(x) = f(x)dx$ where f is a PDF.

nb. (i) $\Longrightarrow$ It takes into account the whole distribution (i.e. "all moments")

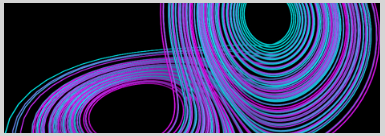


(ii) It is versatile.

(ii) It is versatile.

You can compare *any* two probability distributions:

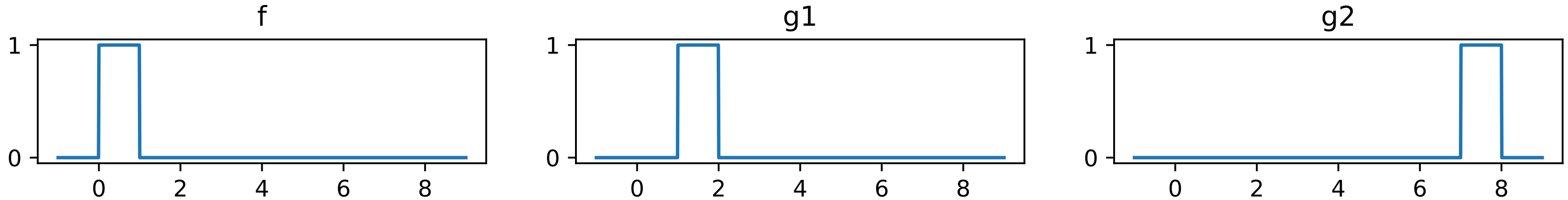
- Continuous distributions.
- Discrete / singular distributions.
- Distributions defined on different spaces.



(iii) It respects the geometry of the underlying space.

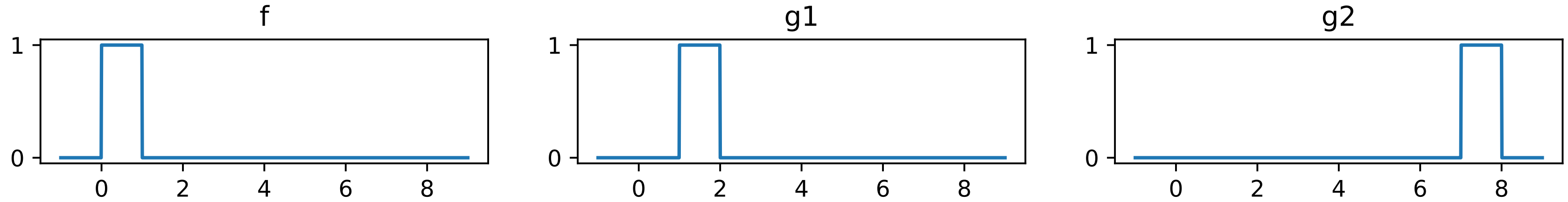
(iii) It respects the geometry of the underlying space.

- Consider the following 3 simple PDFs:



(iii) It respects the geometry of the underlying space.

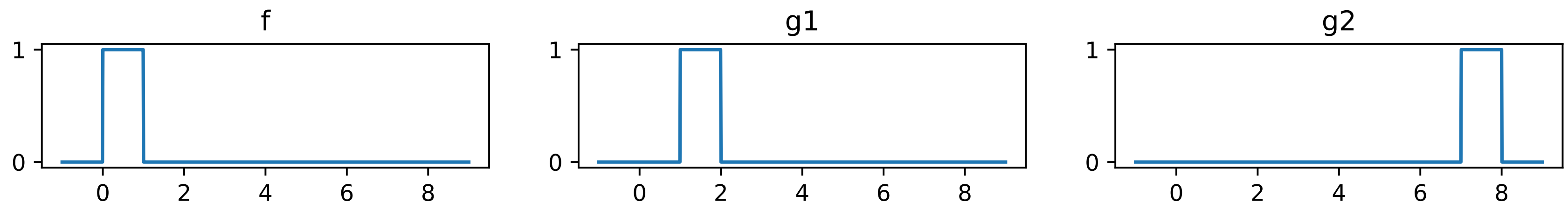
- Consider the following 3 simple PDFs:



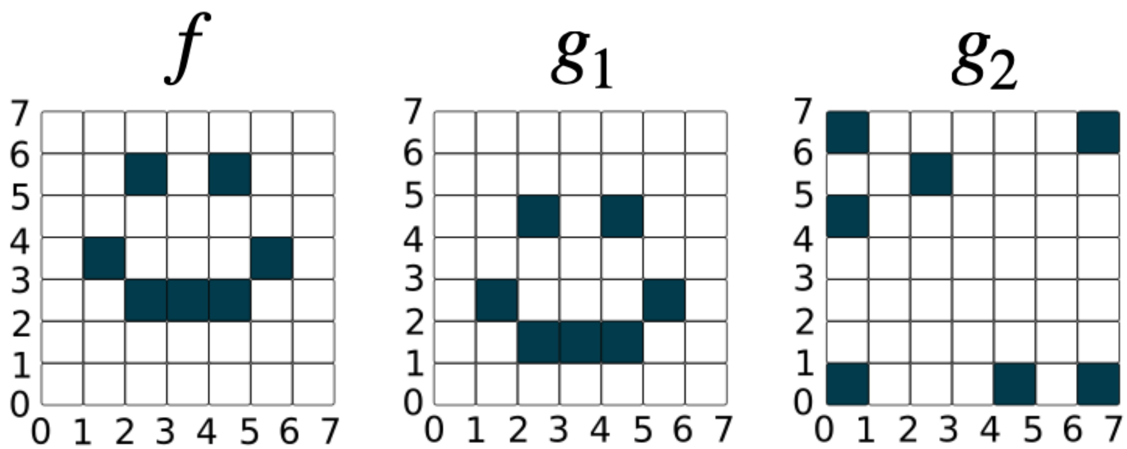
With L^p-distance we have $\|f - g_1\|_{L^p} = \|f - g_2\|_{L^p} = 2$.
But $\mathrm{WD}_1(f, g_1) = 1$, $\mathrm{WD}_1(f, g_2) = 7$.

(iii) It respects the geometry of the underlying space.

- Consider the following 3 simple PDFs:



With L^p-distance we have $\|f - g_1\|_{L^p} = \|f - g_2\|_{L^p} = 2$.
But $\mathrm{WD}_1(f, g_1) = 1$, $\mathrm{WD}_1(f, g_2) = 7$.

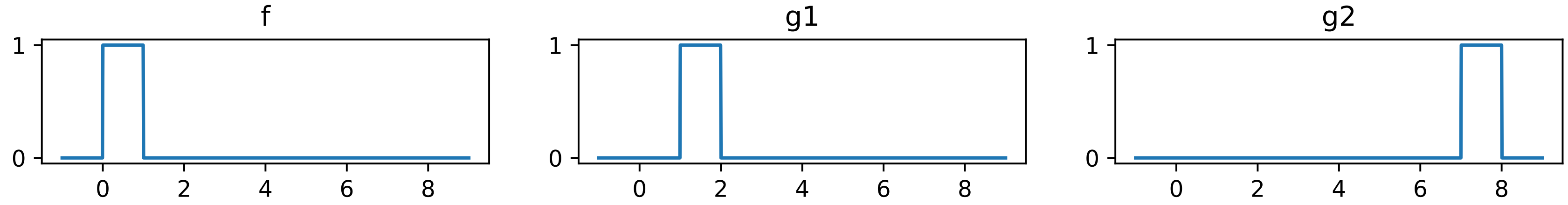


- This is only worse in higher dimension!
Here we have $\|f - g_1\|_{L^p} > \|f - g_2\|_{L^p}$ while $\mathrm{WD}_1(f, g_1) < \mathrm{WD}_1(f, g_2)$.

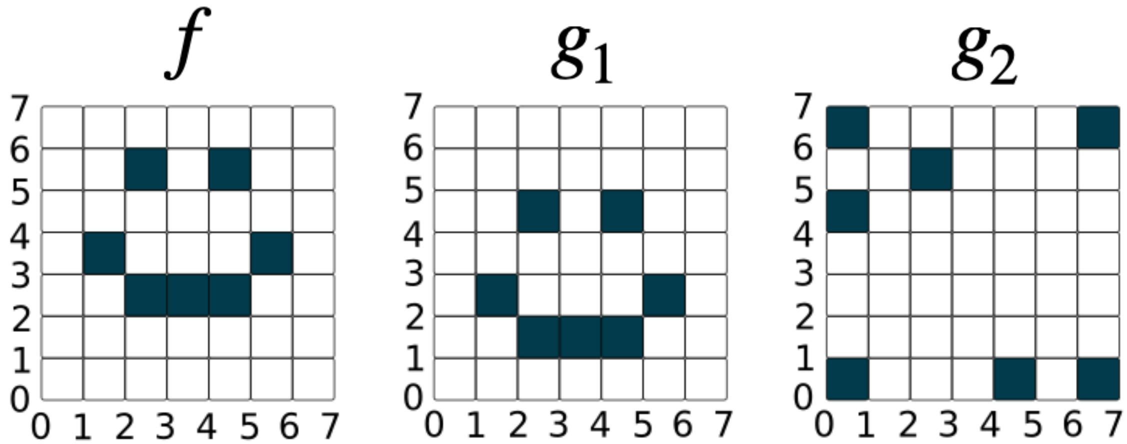
(iii) It respects the geometry of the underlying space.

Nb. This is a shortcoming of many common metrics e.g. K-S test / K-L divergence

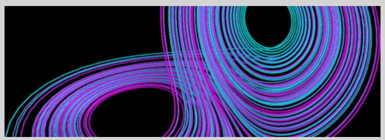
- Consider the following 3 simple PDFs:



With L^p-distance we have $\|f - g_1\|_{L^p} = \|f - g_2\|_{L^p} = 2$.
But $\mathrm{WD}_1(f, g_1) = 1$, $\mathrm{WD}_1(f, g_2) = 7$.



- This is only worse in higher dimension!
Here we have $\|f - g_1\|_{L^p} > \|f - g_2\|_{L^p}$ while $\mathrm{WD}_1(f, g_1) < \mathrm{WD}_1(f, g_2)$.



Computation of the WD:

Computation of the WD:

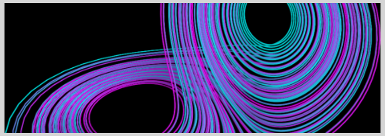
- Monge formulation: $$\mathrm{WD}(\mu,\nu) = \min_{\sigma \in S_N} \frac{1}{N} \sum_{i=1}^{N} c(x_i, y_{\sigma(i)})$$
  - Special case of *assignment problem*: “given N workers and N jobs, find the optimal assignment of workers to jobs”.

Computation of the WD:

- Monge formulation: $$\mathrm{WD}(\mu,\nu) = \min_{\sigma \in S_N} \frac{1}{N} \sum_{i=1}^{N} c(x_i, y_{\sigma(i)})$$
  - Special case of *assignment problem*: “given N workers and N jobs, find the optimal assignment of workers to jobs”.
  - Can be solved in $\mathcal{O}(N^3)$ with Hungarian Algorithm (actually discovered by Jacobi).

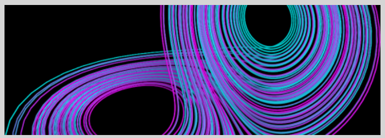
Computation of the WD:

- Monge formulation: $$\mathrm{WD}(\mu,\nu) = \min_{\sigma \in S_N} \frac{1}{N} \sum_{i=1}^{N} c(x_i, y_{\sigma(i)})$$
  - Special case of *assignment problem*: “given N workers and N jobs, find the optimal assignment of workers to jobs”.
  - Can be solved in $\mathcal{O}(N^3)$ with Hungarian Algorithm (actually discovered by Jacobi).
- Kantorovich formulation: $$\mathrm{WD}(\mu,\nu) = \min_{\pi \in \Pi(\mu,\nu)} \sum_{i,j=1}^{M_1,M_2} c_{ij}\pi_{ij}$$
  - Case of *linear programming*. Can (usually) be solved in polynomial time by e.g simplex algorithm.



Computation of the WD:

- Monge formulation: $$\mathrm{WD}(\mu,\nu) = \min_{\sigma\in S_N} \frac{1}{N}\sum_{i=1}^{N} c(x_i, y_{\sigma(i)})$$
  - Special case of *assignment problem*: “given N workers and N jobs, find the optimal assignment of workers to jobs”.
  - Can be solved in $\mathcal{O}(N^3)$ with Hungarian Algorithm (actually discovered by Jacobi).
- Kantorovich formulation: $$\mathrm{WD}(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \sum_{i,j=1}^{M_1,M_2} c_{ij}\pi_{ij}$$
  - Case of *linear programming*. Can (usually) be solved in polynomial time by e.g simplex algorithm.
- Approximate formulations (e.g. Cuturi: *Sinkhorn Distances: Lightspeed Computation of Optimal Transport*)



Computation of the WD:

- Monge formulation: $$\mathrm{WD}(\mu,\nu) = \min_{\sigma \in S_N} \frac{1}{N} \sum_{i=1}^{N} c(x_i, y_{\sigma(i)})$$

Nb. This scales with N = number of *samples*.

  - Special case of *assignment problem*: "given N workers and N jobs, find the optimal assignment of workers to jobs".

  - Can be solved in $\mathcal{O}(N^3)$ with Hungarian Algorithm (actually discovered by Jacobi).

- Kantorovich formulation: $$\mathrm{WD}(\mu,\nu) = \min_{\pi \in \Pi(\mu,\nu)} \sum_{i,j=1}^{M_1,M_2} c_{ij}\pi_{ij}$$

This scales with M = number of *bins*.

  - Case of *linear programming*. Can (usually) be solved in polynomial time by e.g simplex algorithm.

- Approximate formulations (e.g. Cuturi: *Sinkhorn Distances: Lightspeed Computation of Optimal Transport*)

Computation of the WD:

- Monge formulation: $$\mathrm{WD}(\mu,\nu) = \min_{\sigma\in S_N} \frac{1}{N}\sum_{i=1}^{N} c(x_i, y_{\sigma(i)})$$

Nb. This scales with N = number of *samples*.

  - Special case of *assignment problem*: “given N workers and N jobs, find the optimal assignment of workers to jobs”.
  - Can be solved in $\mathcal{O}(N^3)$ with Hungarian Algorithm (actually discovered by Jacobi).

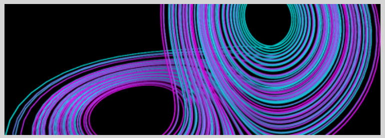
- Kantorovich formulation: $$\mathrm{WD}(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \sum_{i,j=1}^{M_1,M_2} c_{ij}\pi_{ij}$$

This scales with M = number of *bins*.

  - Case of *linear programming*. Can (usually) be solved in polynomial time by e.g simplex algorithm.

- Approximate formulations (e.g. Cuturi: *Sinkhorn Distances: Lightspeed Computation of Optimal Transport*)

All of these can be found at github.com/eapax/EarthMover.jl



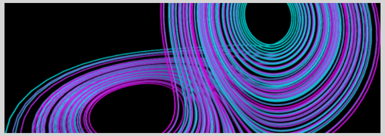
3) An application: exploring model climatology in low-precision.

3) An application: exploring model climatology in low-precision.

- Recently there has been lots of interest in low (<64bit) precision arithmetic for high-performance computing.

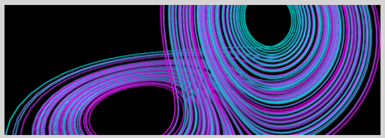
3) An application: exploring model climatology in low-precision.

- Recently there has been lots of interest in low (<64bit) precision arithmetic for high-performance computing.
- Operational weather forecasting centres have begun porting models to low-precision.



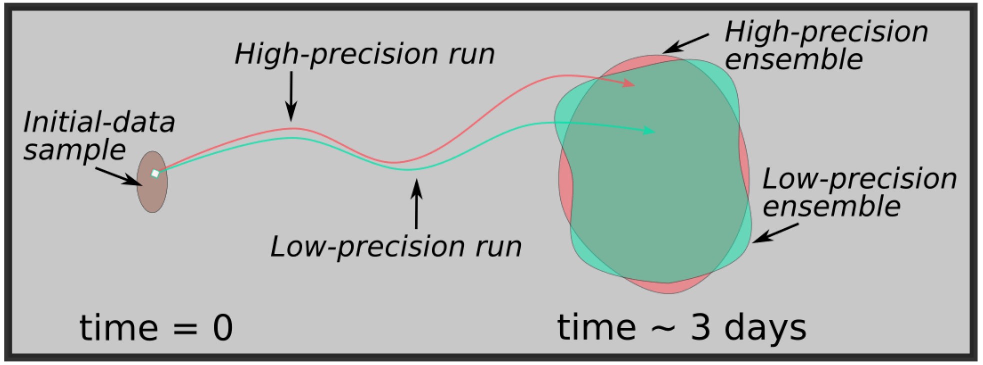
3) An application: exploring model climatology in low-precision.

- Recently there has been lots of interest in low (<64bit) precision arithmetic for high-performance computing.
- Operational weather forecasting centres have begun porting models to low-precision.
- As forecast models move to low-precision, it's natural to ask if these models are suitable for climate modelling (some have argued NOT).



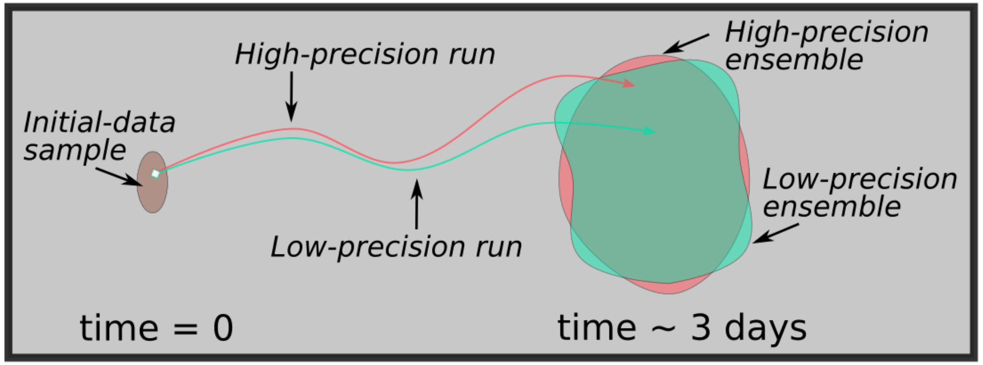
Climate modelling & weather forecasting are different methodologies.

Test for low-precision weather forecast	Test for low-precision climate model
Does it produce the same probabilistic ensemble forecast as high-precision?	*?*



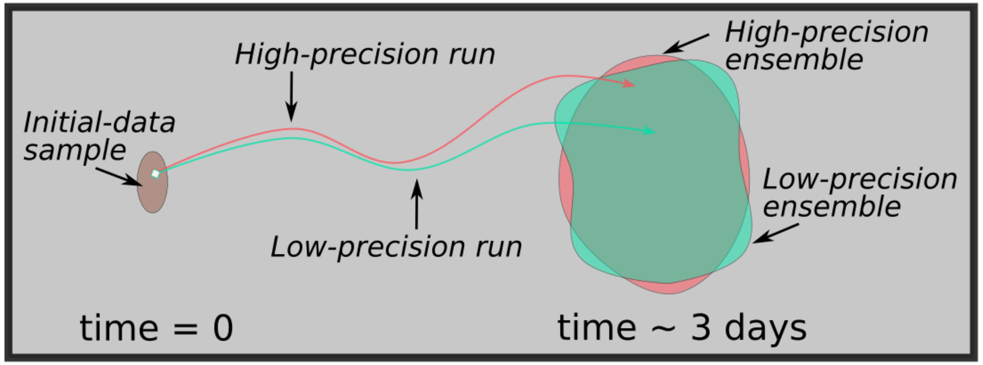
Climate modelling & weather forecasting are different methodologies.

Test for low-precision weather forecast	Test for low-precision climate model
Does it produce the same probabilistic ensemble forecast as high-precision?	*Does it produce the same long-time statistics (invariant measure) as high-precision?*

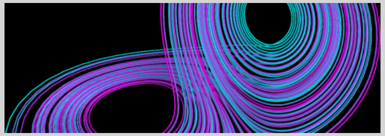


Climate modelling & weather forecasting are different methodologies.

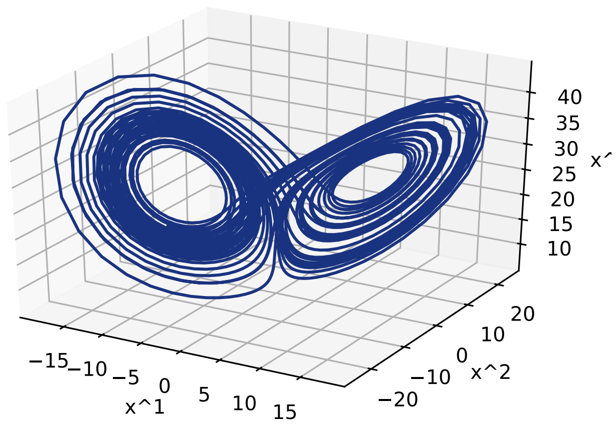
Test for low-precision weather forecast	Test for low-precision climate model
Does it produce the same probabilistic ensemble forecast as high-precision?	*Does it produce the same long-time statistics (invariant measure) as high-precision?*



Idea: use the Wasserstein Distance to test this.



Example: L63 (toy model).



$$x(t) = (x^1(t), x^2(t), x^3(t));$$

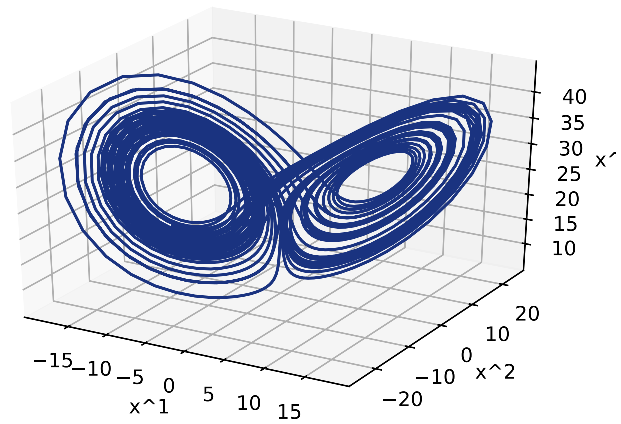
$$\dot{x}^1 = 10\left(x^2 - x^1\right)$$

$$\dot{x}^2 = \left(\frac{8}{3} - x^3\right) x^1 - x^2$$

$$\dot{x}^3 = x^1 x^2 - 28 x^3$$

Example: L63 (toy model).

- Admits an attractor $\mathscr{A} \subseteq \mathbb{R}^3$ $(x(t) \to \mathscr{A}$ as $t \to \infty)$.

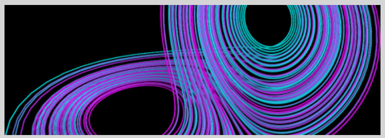


$$x(t) = (x^1(t), x^2(t), x^3(t));$$

$$\dot{x}^1 = 10\left(x^2 - x^1\right)$$

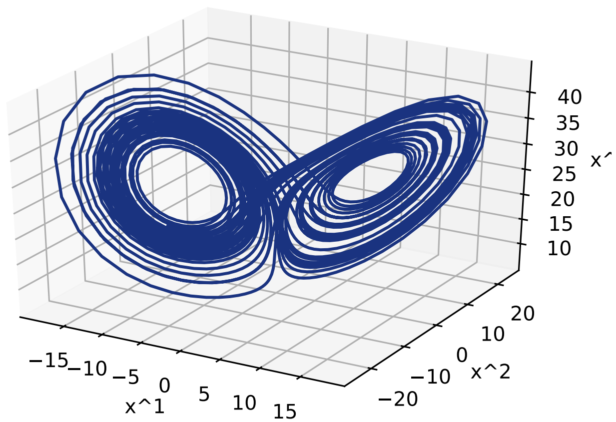
$$\dot{x}^2 = \left(\frac{8}{3} - x^3\right)x^1 - x^2$$

$$\dot{x}^3 = x^1x^2 - 28x^3$$



Example: L63 (toy model).

- Admits an attractor $\mathscr{A} \subseteq \mathbb{R}^3$ $(x(t) \to \mathscr{A}$ as $t \to \infty)$.
- $\mathscr{A}$ is chaotic (positive Lyapunov exponent).



$$x(t) = (x^1(t), x^2(t), x^3(t));$$

$$\dot{x}^1 = 10\left(x^2 - x^1\right)$$

$$\dot{x}^2 = \left(\frac{8}{3} - x^3\right)x^1 - x^2$$

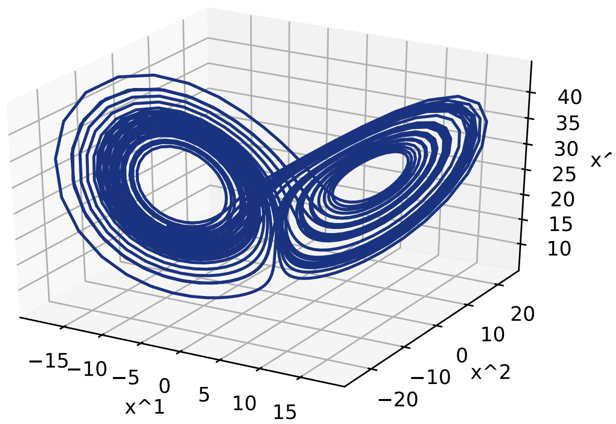
$$\dot{x}^3 = x^1x^2 - 28x^3$$

Example: L63 (toy model).

- Admits an attractor $\mathscr{A} \subseteq \mathbb{R}^3$ $(x(t) \to \mathscr{A}$ as $t \to \infty)$.
- $\mathscr{A}$ is chaotic (positive Lyapunov exponent).
- Admits an *invariant probability measure* μ supported on $\mathscr{A}$ such that

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \phi(x(t))dt = \iiint_{\mathbb{R}^3} \phi(x)d\mu(x)$$

for any solution $x(t)$ and any bounded function $\phi(x)$.
i.e. μ encodes the long-time statistics of the system.

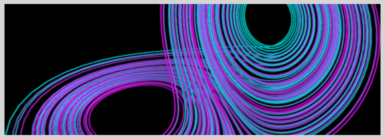


$$x(t) = (x^1(t), x^2(t), x^3(t));$$

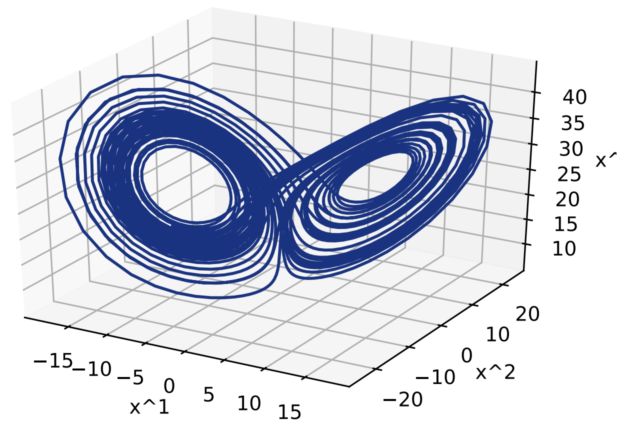
$$\dot{x}^1 = 10\left(x^2 - x^1\right)$$

$$\dot{x}^2 = \left(\frac{8}{3} - x^3\right)x^1 - x^2$$

$$\dot{x}^3 = x^1x^2 - 28x^3$$



Example: L63 (toy model).



- Admits an attractor $\mathscr{A} \subseteq \mathbb{R}^3$ $(x(t) \to \mathscr{A}$ as $t \to \infty)$.
- $\mathscr{A}$ is chaotic (positive Lyapunov exponent).
- Admits an *invariant probability measure* μ supported on $\mathscr{A}$ such that

e.g. take $\phi(x) = \begin{cases} 1 & x \in B \\ 0 & x \notin B \end{cases}$

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \phi(x(t))dt = \iiint_{\mathbb{R}^3} \phi(x)d\mu(x)$$

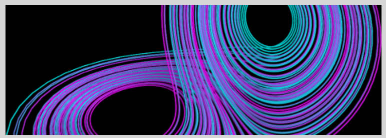
for any solution $x(t)$ and any bounded function $\phi(x)$.
i.e. μ encodes the long-time statistics of the system.

$$x(t) = (x^1(t), x^2(t), x^3(t));$$

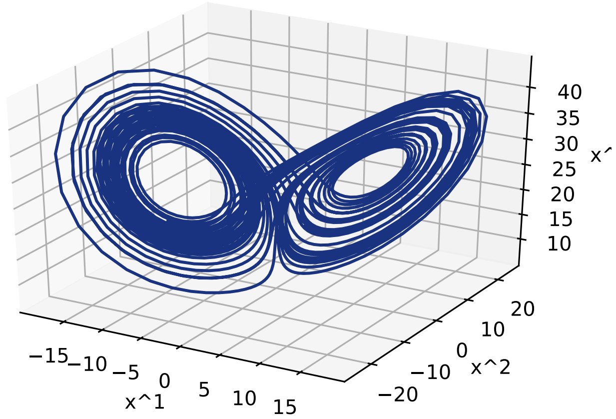
$$\dot{x}^1 = 10\left(x^2 - x^1\right)$$

$$\dot{x}^2 = \left(\frac{8}{3} - x^3\right)x^1 - x^2$$

$$\dot{x}^3 = x^1x^2 - 28x^3$$



Example: L63 (toy model).



- Admits an attractor $\mathscr{A} \subseteq \mathbb{R}^3$ $(x(t) \to \mathscr{A}$ as $t \to \infty)$.
- $\mathscr{A}$ is chaotic (positive Lyapunov exponent).
- Admits an *invariant probability measure* μ supported on $\mathscr{A}$ such that

e.g. take $\phi(x) = \begin{cases} 1 & x \in B \\ 0 & x \notin B \end{cases}$

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \phi(x(t))dt = \iiint_{\mathbb{R}^3} \phi(x)d\mu(x)$$

for any solution $x(t)$ and any bounded function $\phi(x)$.
i.e. μ encodes the long-time statistics of the system.

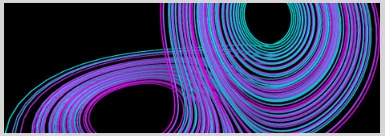
nb. link to weak★ convergence!

$$x(t) = (x^1(t), x^2(t), x^3(t));$$

$$\dot{x}^1 = 10\left(x^2 - x^1\right)$$

$$\dot{x}^2 = \left(\frac{8}{3} - x^3\right)x^1 - x^2$$

$$\dot{x}^3 = x^1x^2 - 28x^3$$



How can we approximate (/visualize) μ?

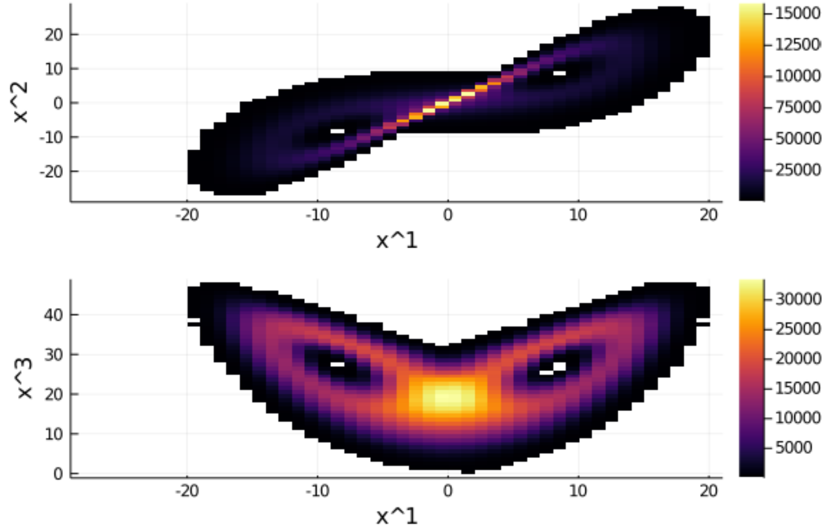
How can we approximate (/visualize) μ?

Two methods:

How can we approximate (/visualize) μ?

Two methods:

1. Data-binning
   (i.e. approximate μ as a histogram)

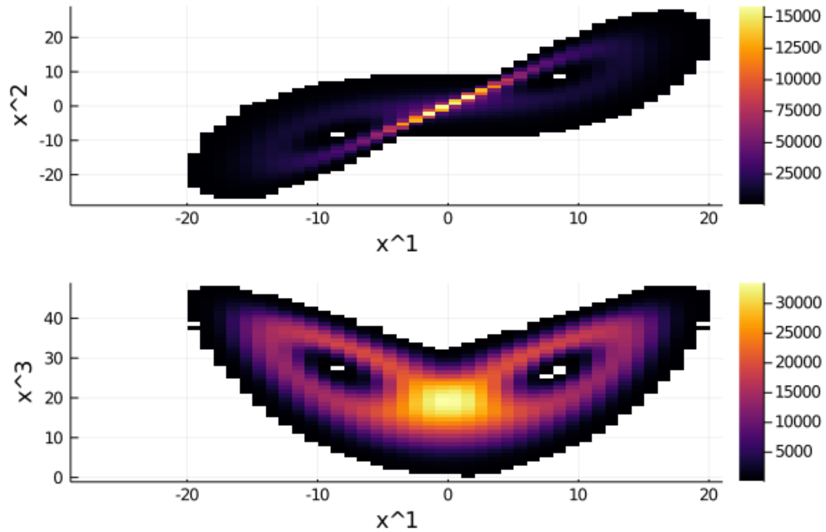


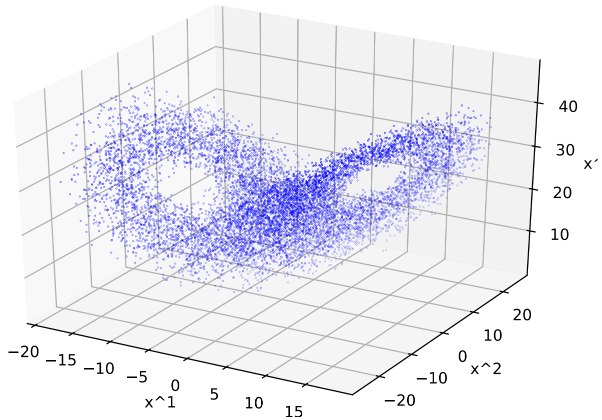
How can we approximate (/visualize) μ?

Two methods:

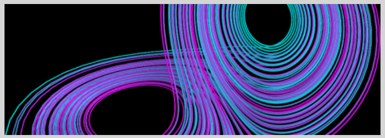
1. Data-binning
   (i.e. approximate μ as a histogram)

2. Scatter-plotting
   (i.e. approximate directly from sampling as $\mu \approx \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$)





E. Adam Paxton

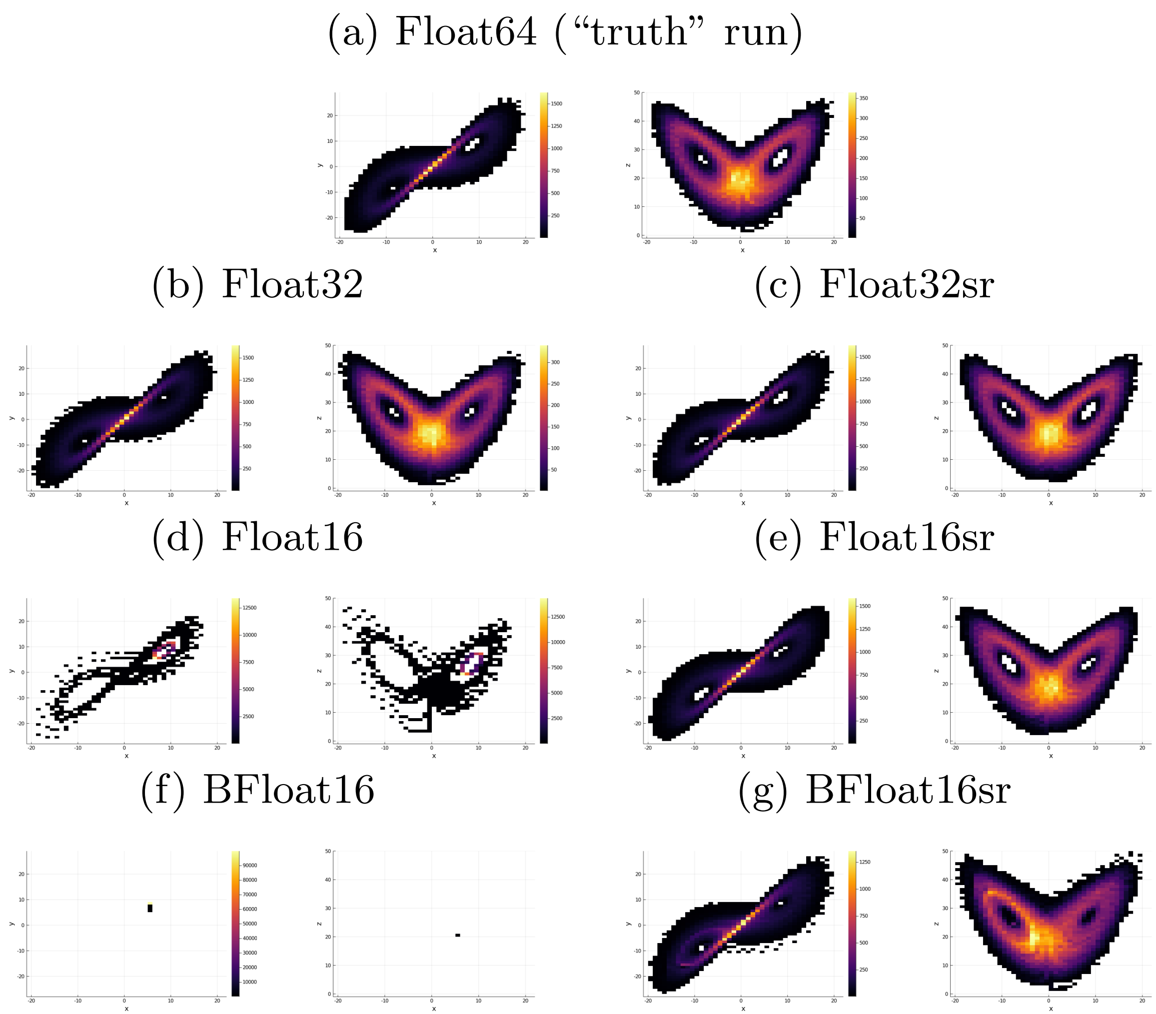


UNIVERSITY OF OXFORD

Now for the reduced precision...

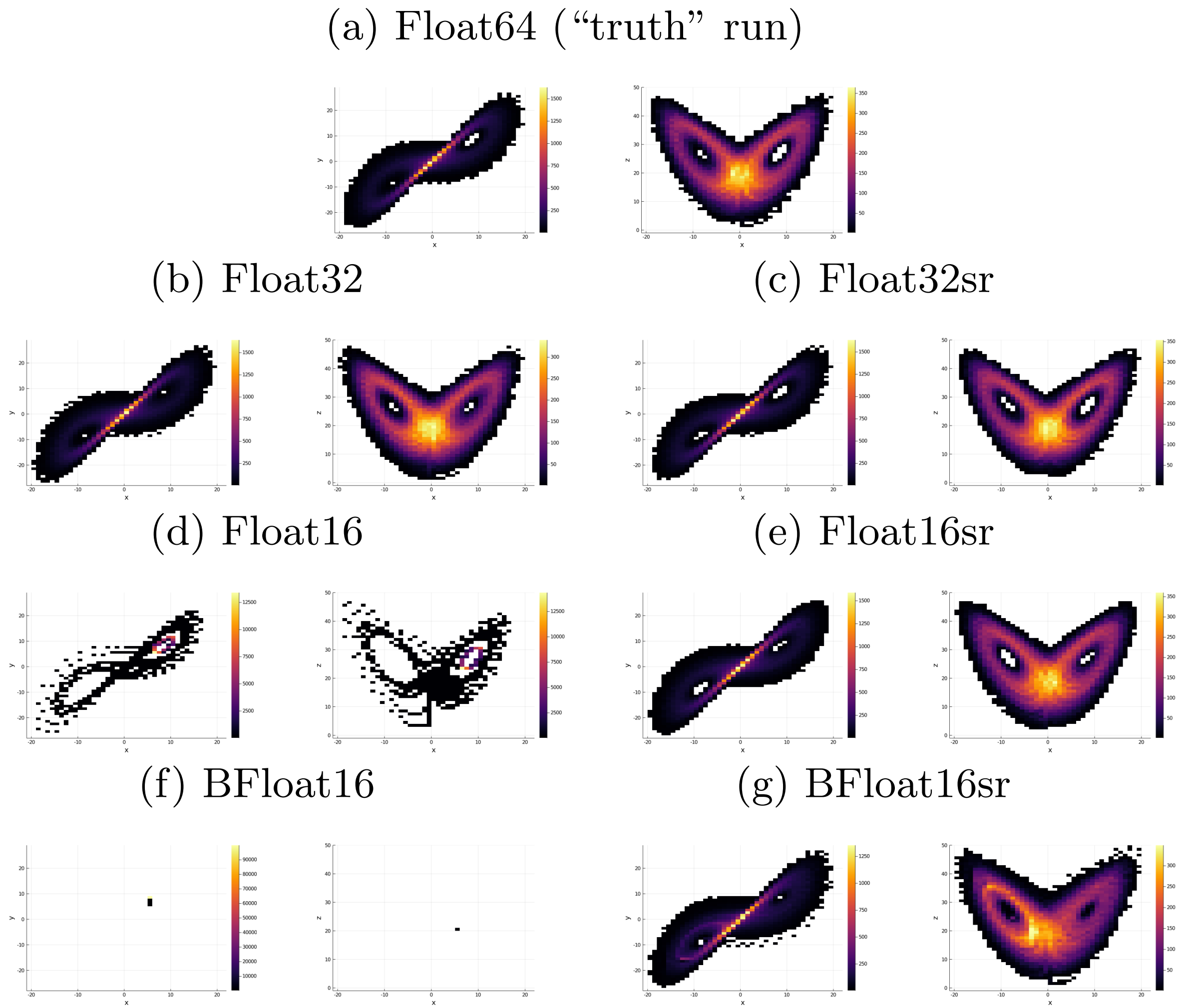
Now for the reduced precision...

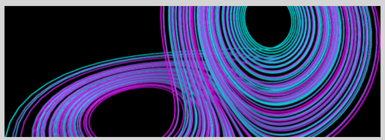
- Integrated L63 in different numerical precisions.



Now for the reduced precision...

- Integrated L63 in different numerical precisions.
- Approximated invariant measures by data-binning.

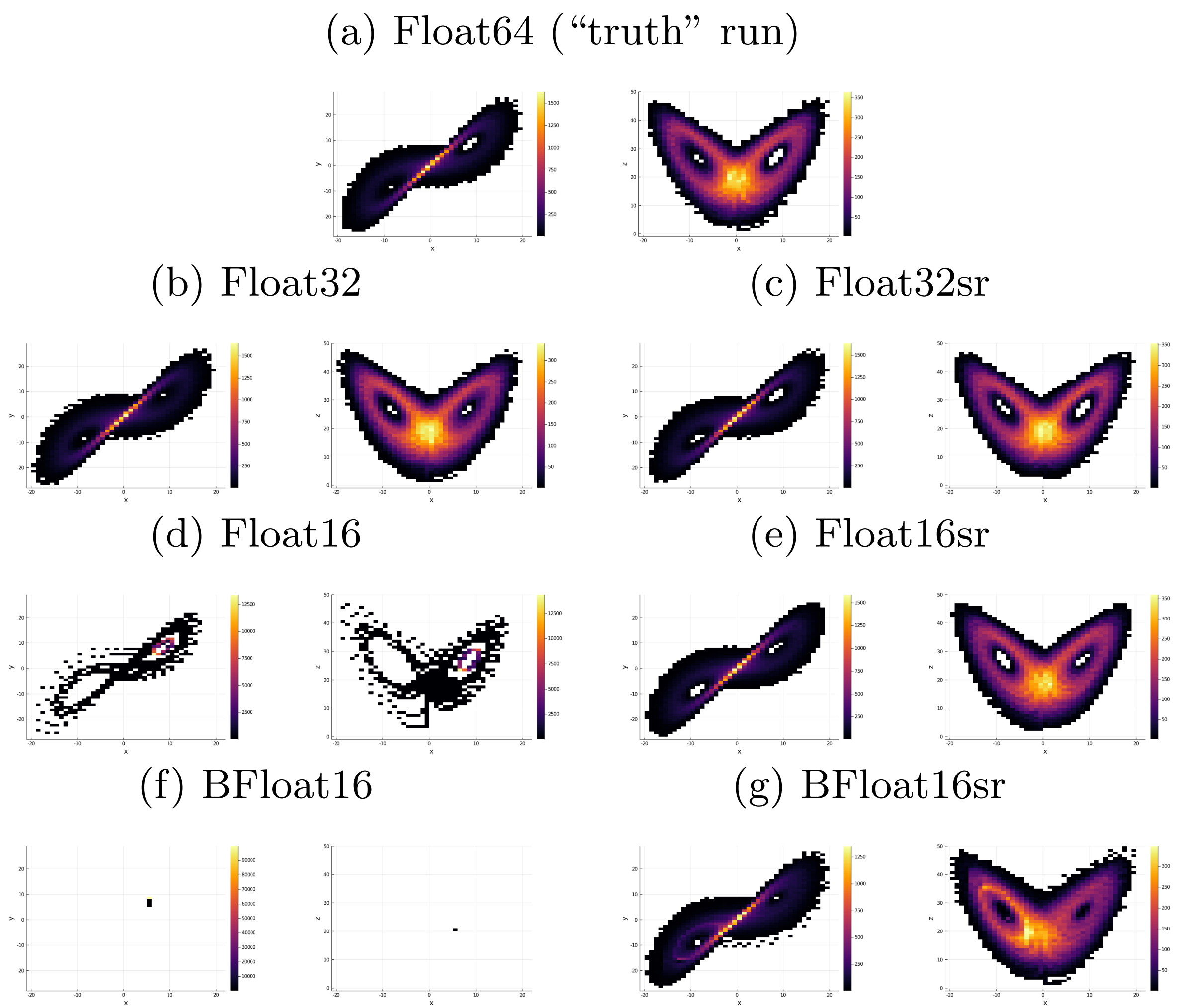


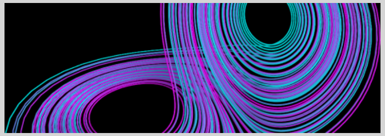


UNIVERSITY OF OXFORD

Now for the reduced precision...

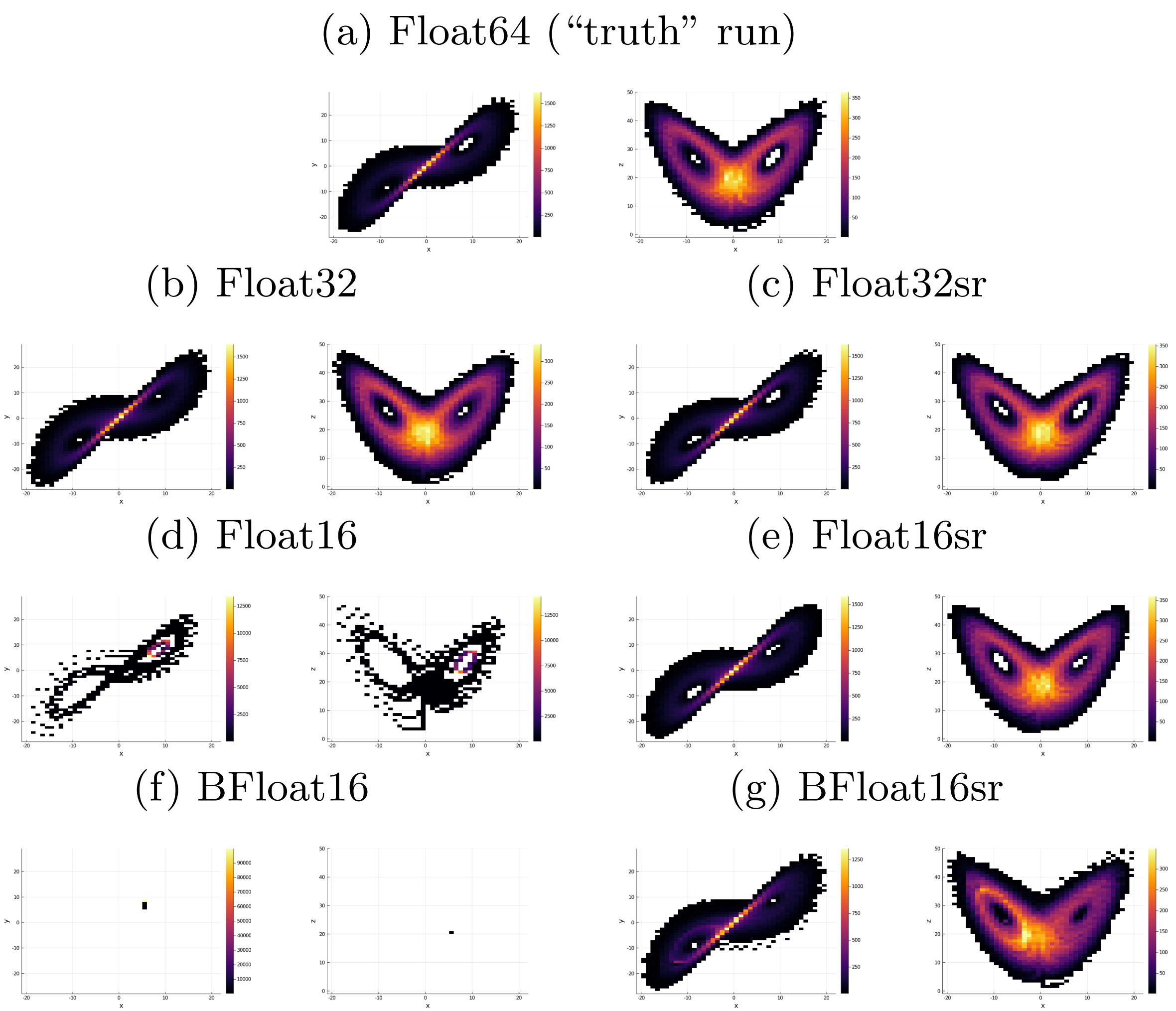
- Integrated L63 in different numerical precisions.
- Approximated invariant measures by data-binning.
- We want a method for quantitative comparison.





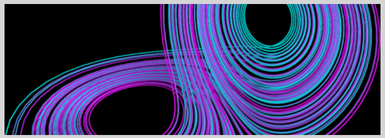
Now for the reduced precision...

- Integrated L63 in different numerical precisions.
- Approximated invariant measures by data-binning.
- We want a method for quantitative comparison.
- Let’s compute the Wasserstein Distances!



- Here are the results...

precision	WD(precision, Float64)
Float64	0.0
Float32	0.456
Float32sr	0.353
Float16	14.8
Float16sr	0.421
BFloat16	16.1
BFloat16sr	3.82

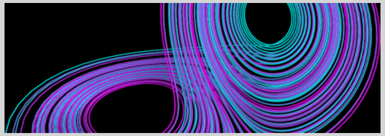


UNIVERSITY OF OXFORD

- Here are the results…

… but what do these numbers mean?

precision	WD(precision, Float64)
Float64	0.0
Float32	0.456
Float32sr	0.353
Float16	14.8
Float16sr	0.421
BFloat16	16.1
BFloat16sr	3.82

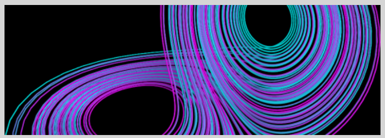


- Here are the results…

… but what do these numbers mean?

- We need a null hypothesis.

precision	WD(precision, Float64)
Float64	0.0
Float32	0.456
Float32sr	0.353
Float16	14.8
Float16sr	0.421
BFloat16	16.1
BFloat16sr	3.82

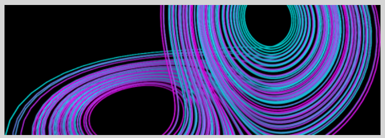


- Here are the results...

  ... but what do these numbers mean?

- We need a null hypothesis.
- Idea: use an *ensemble.*

precision	WD(precision, Float64)
Float64	0.0
Float32	0.456
Float32sr	0.353
Float16	14.8
Float16sr	0.421
BFloat16	16.1
BFloat16sr	3.82



UNIVERSITY OF OXFORD

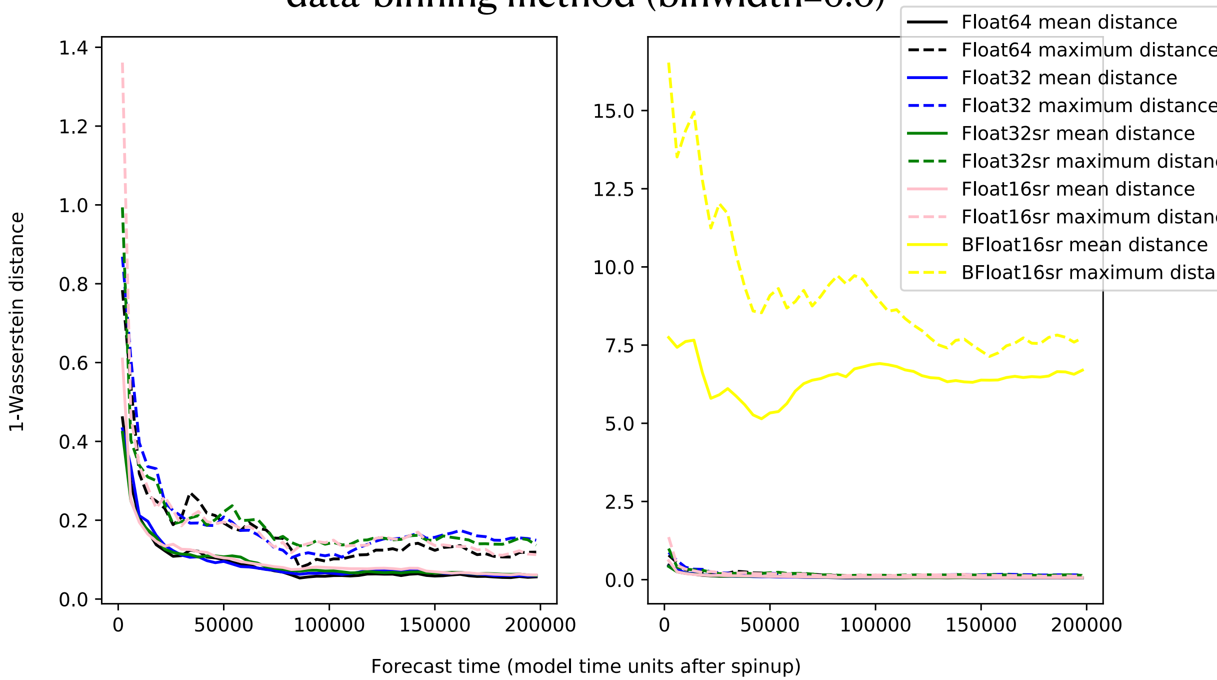
Experiment set-up:

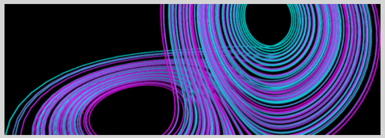
- Take one 5-member Float64 ensemble (Control)
- Take a 5-member ensemble for each precision (including Float64) and compare with the Control pairwise (25 comparisons).
- Plot the mean & maximum values with time.

Experiment set-up:

- Take one 5-member Float64 ensemble (Control)
- Take a 5-member ensemble for each precision (including Float64) and compare with the Control pairwise (25 comparisons).
- Plot the mean & maximum values with time.

Convergence to statistical equilibrium:
data-binning method (binwidth=6.0)

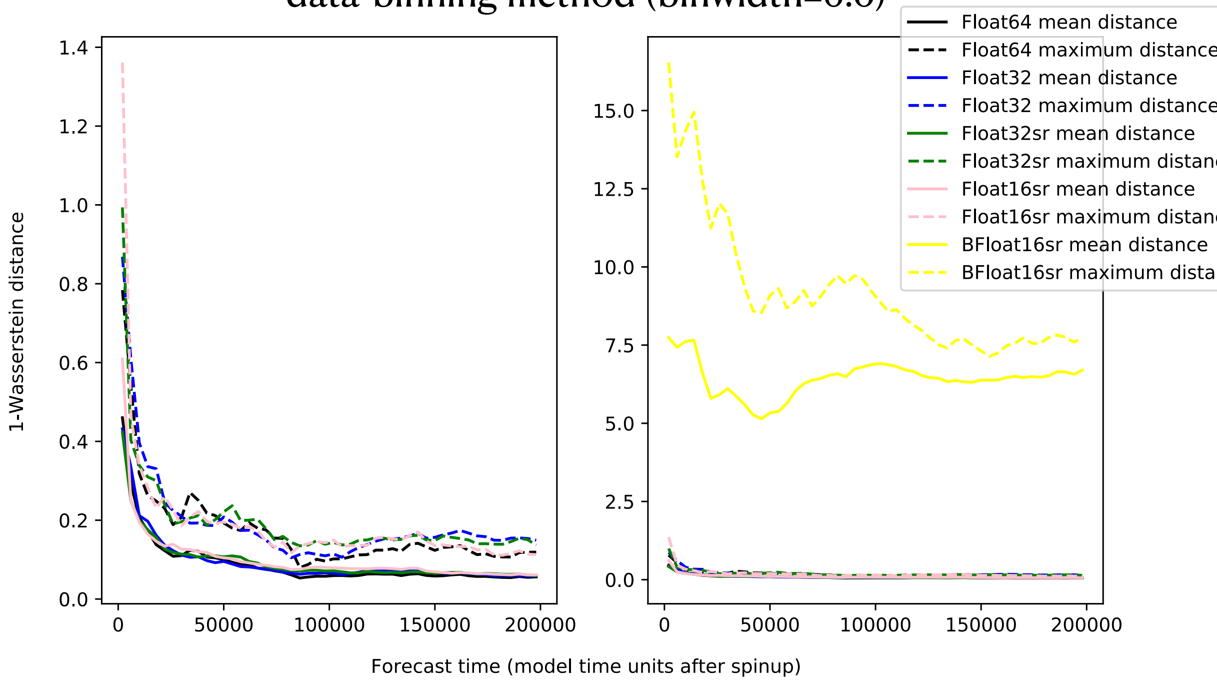




Experiment set-up:

- Take one 5-member Float64 ensemble (Control)
- Take a 5-member ensemble for each precision (including Float64) and compare with the Control pairwise (25 comparisons).
- Plot the mean & maximum values with time.

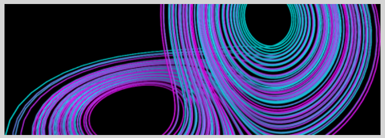
Convergence to statistical equilibrium:
data-binning method (binwidth=6.0)



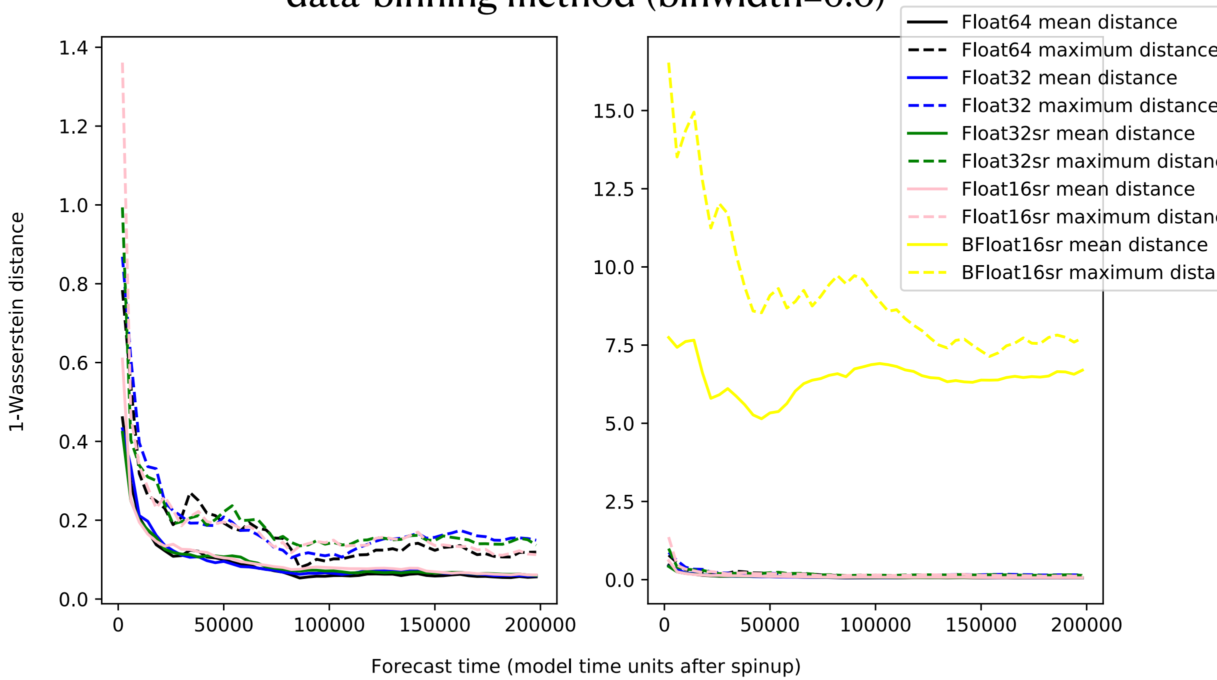
The Float64 vs Control test (black lines) serves 2 purposes:

1. It gives a null hypothesis.

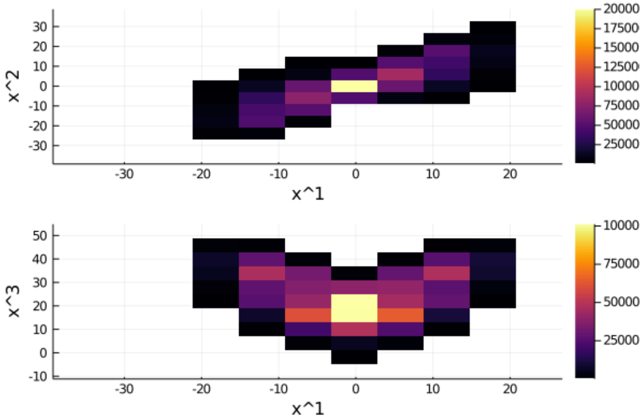
2. It shows that enough time has elapsed to reach statistical equilibrium.

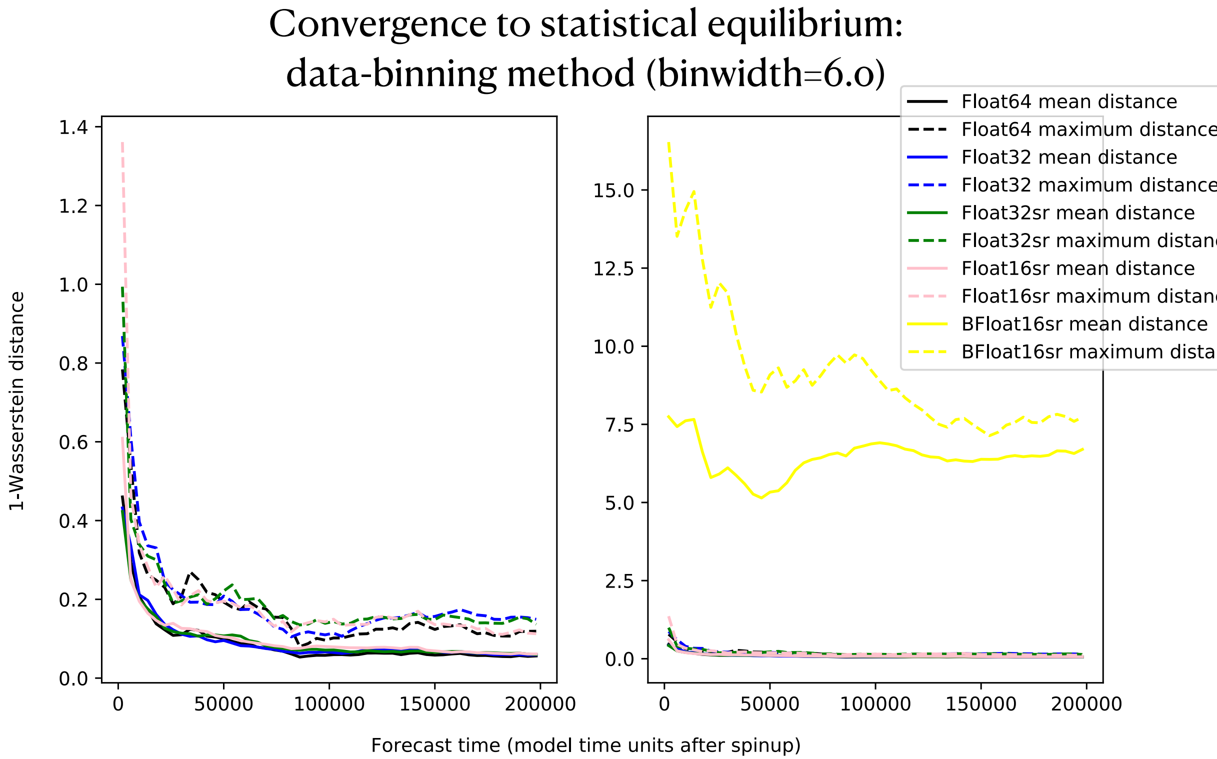


Convergence to statistical equilibrium: data-binning method (binwidth=6.0)

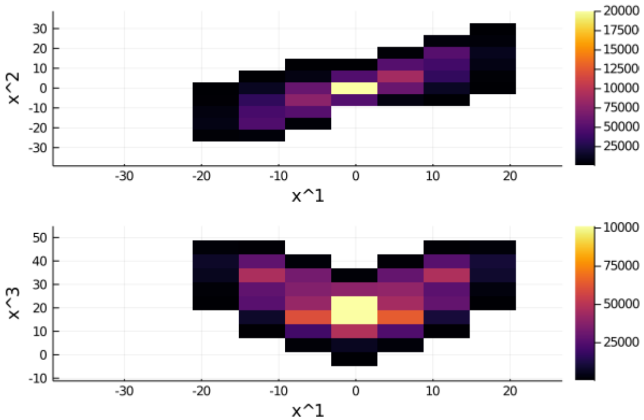


nb. bin-width=6.0 looks like:

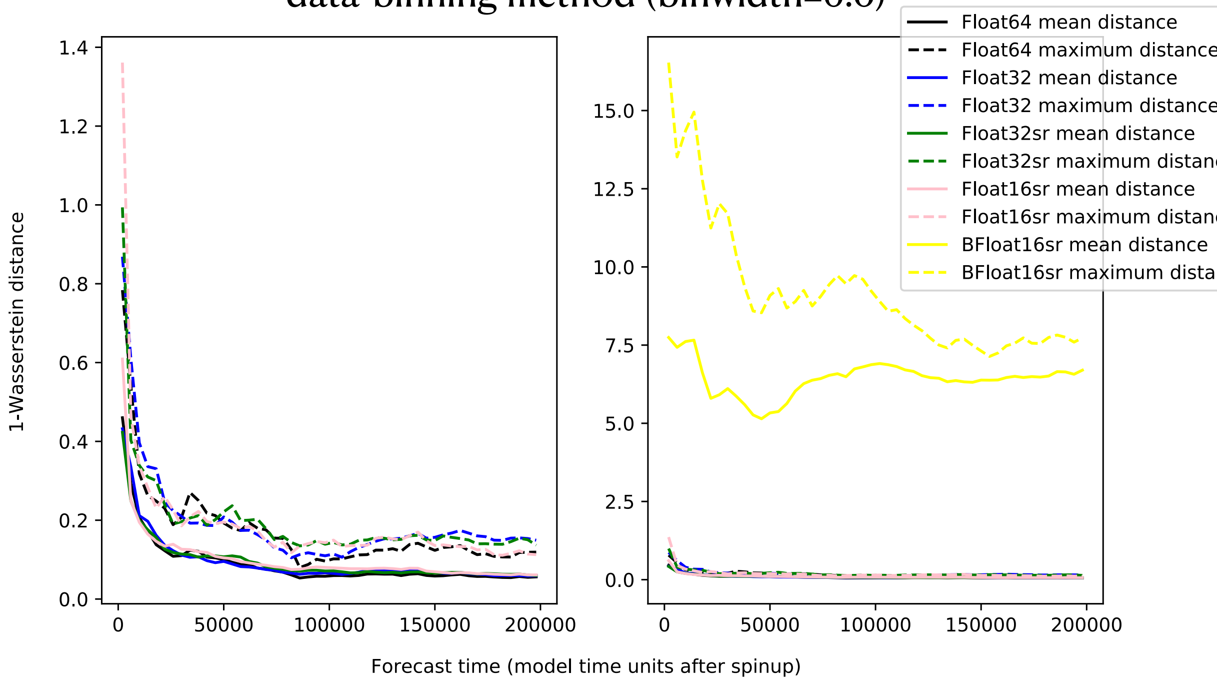




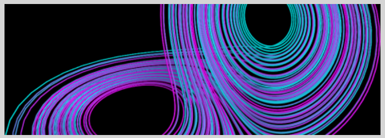
nb. bin-width=6.0 looks like:



Convergence to statistical equilibrium:
data-binning method (binwidth=6.0)



- Results are not sensitive to decreasing bin-width.

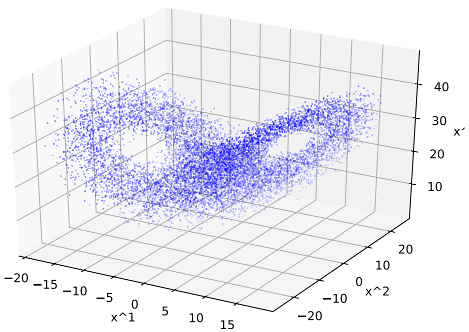


UNIVERSITY OF OXFORD

Note: the "scatter-plot method" is also available.

Note: the "scatter-plot method" is also available

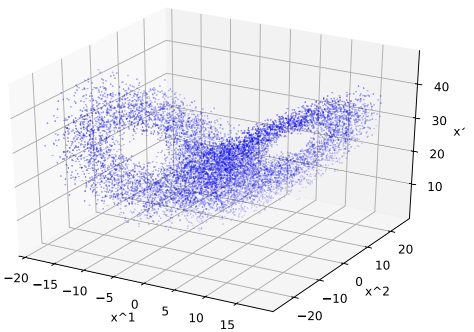
(i.e. approximate as
$\mu \approx \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$).



Note: the "scatter-plot method" is also available

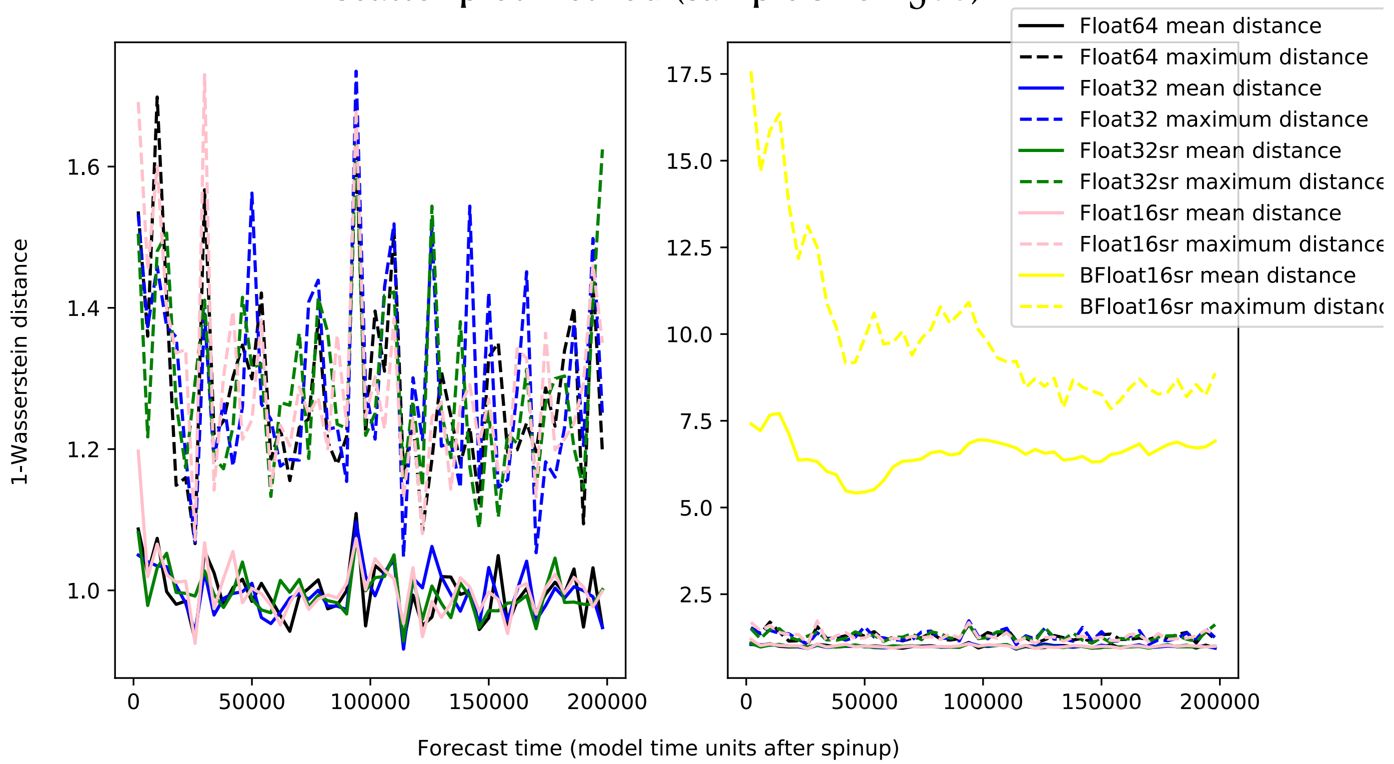
(i.e. approximate as

$$\mu \approx \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}).$$



It gives comparable results.

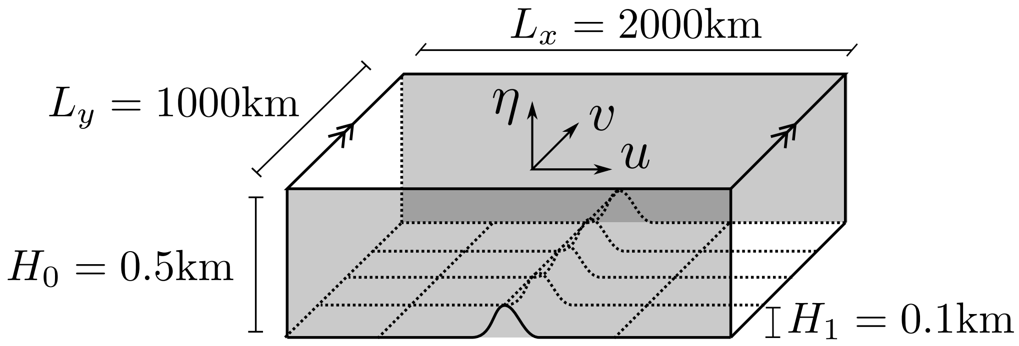
Convergence to statistical equilibrium:
scatter-plot method (sample size=2500)



Shallow Water Model: github.com/milankl/ShallowWaters.jl

Shallow Water Model:

github.com/milankl/ShallowWaters.jl



$\mathbf{u}(x, y, t) = (u(x, y, t), v(x, y, t))$ fluid velocity

$h(x, y, t) = H(x) + \eta(x, y, t)$ layer depth

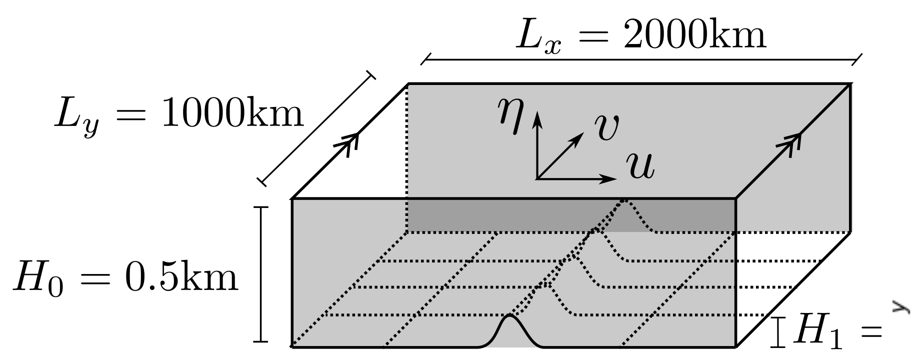
$\mathbf{F}(x, y, t) = (f(y), 0)$ wind forcing

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\,\mathbf{u} + f\mathbf{z} \times \mathbf{u} = -g\nabla\mathbf{u} + \mathbf{D}(\mathbf{u}, \nabla\mathbf{u}) + \mathbf{F}$$

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot (h\mathbf{u}) = 0$$

Shallow Water Model:

github.com/milankl/ShallowWaters.jl



$\mathbf{u}(x,y,t) = (u(x,y,t), v(x,y,t))$ fluid velocity

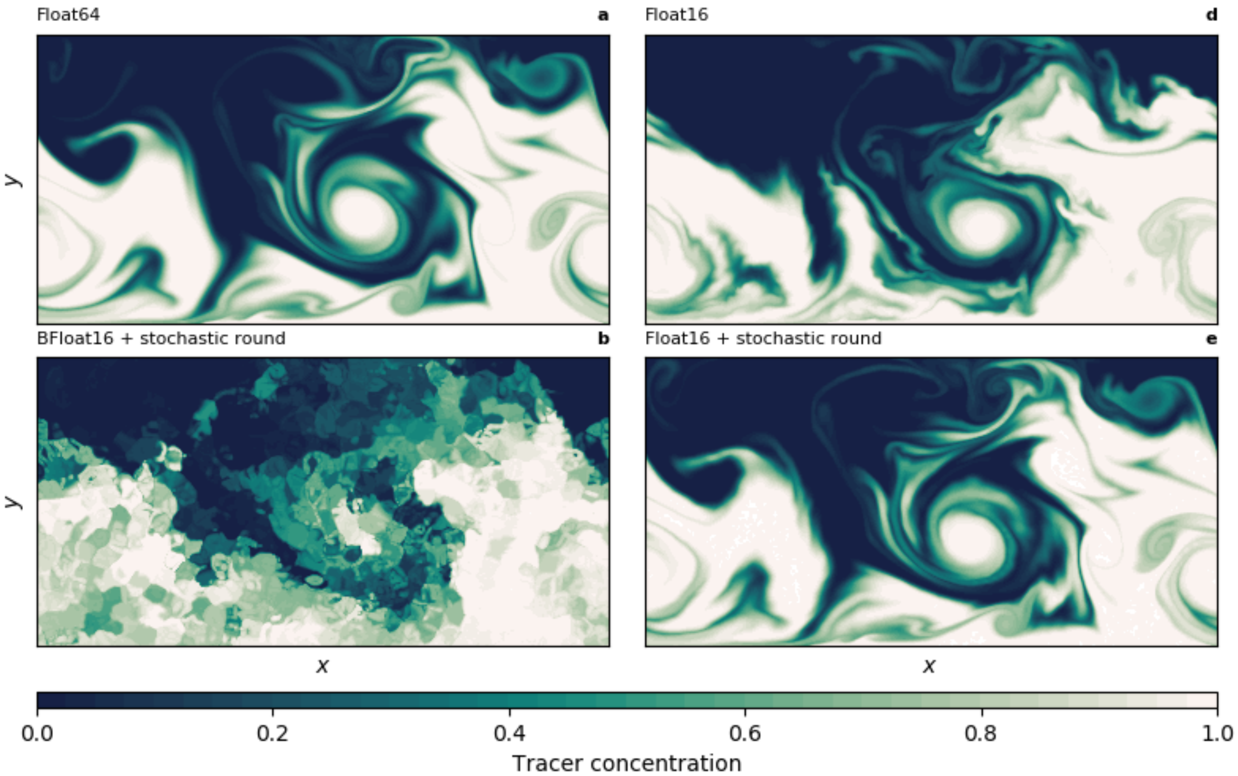
$h(x,y,t) = H(x) + \eta(x,y,t)$ layer depth

$\mathbf{F}(x,y,t) = (f(y), 0)$ wind forcing

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\,\mathbf{u} + f\mathbf{z} \times \mathbf{u} = -g\nabla \mathbf{u} + \mathbf{D}(\mathbf{u}, \nabla \mathbf{u}) + \mathbf{F}$$

$$\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot (h\mathbf{u}) = 0$$

- Finite difference scheme, 100×50 spatial grid

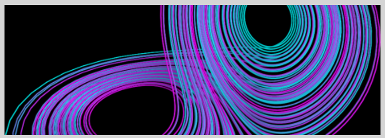


We want to estimate the Shallow Water model climatology (i.e. invariant measure).

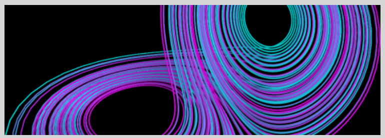
We want to estimate the Shallow Water model climatology (i.e. invariant measure).

Some problems arise:

- We have time evolution in a $100 \times 50 = 5000$ dimensional space.
- Working with high-dimensional probability distributions is non-trivial.
- Data-binning becomes stupid. Looking at just one parameter u and assigning just 2 bins per spatial coordinate would lead to 2^{5000} bins. (number of atoms in observable universe $\approx 2^{270}$)



- One strategy: project down onto lower-dimensional subspaces.



- One strategy: project down onto lower-dimensional subspaces.
- This is what I have seen done so far.

Ranking IPCC Models Using the Wasserstein Distance

G. Vissio[1], V. Lembo[1], V. Lucarini[1,2,3] and M. Ghil[4,5]

[1]CEN, Meteorological Institute, University of Hamburg, Hamburg, Germany

[2]Department of Mathematics and Statistics, University of Reading, Reading, UK

[3]Centre for the Mathematics of Planet Earth, University of Reading, Reading, UK

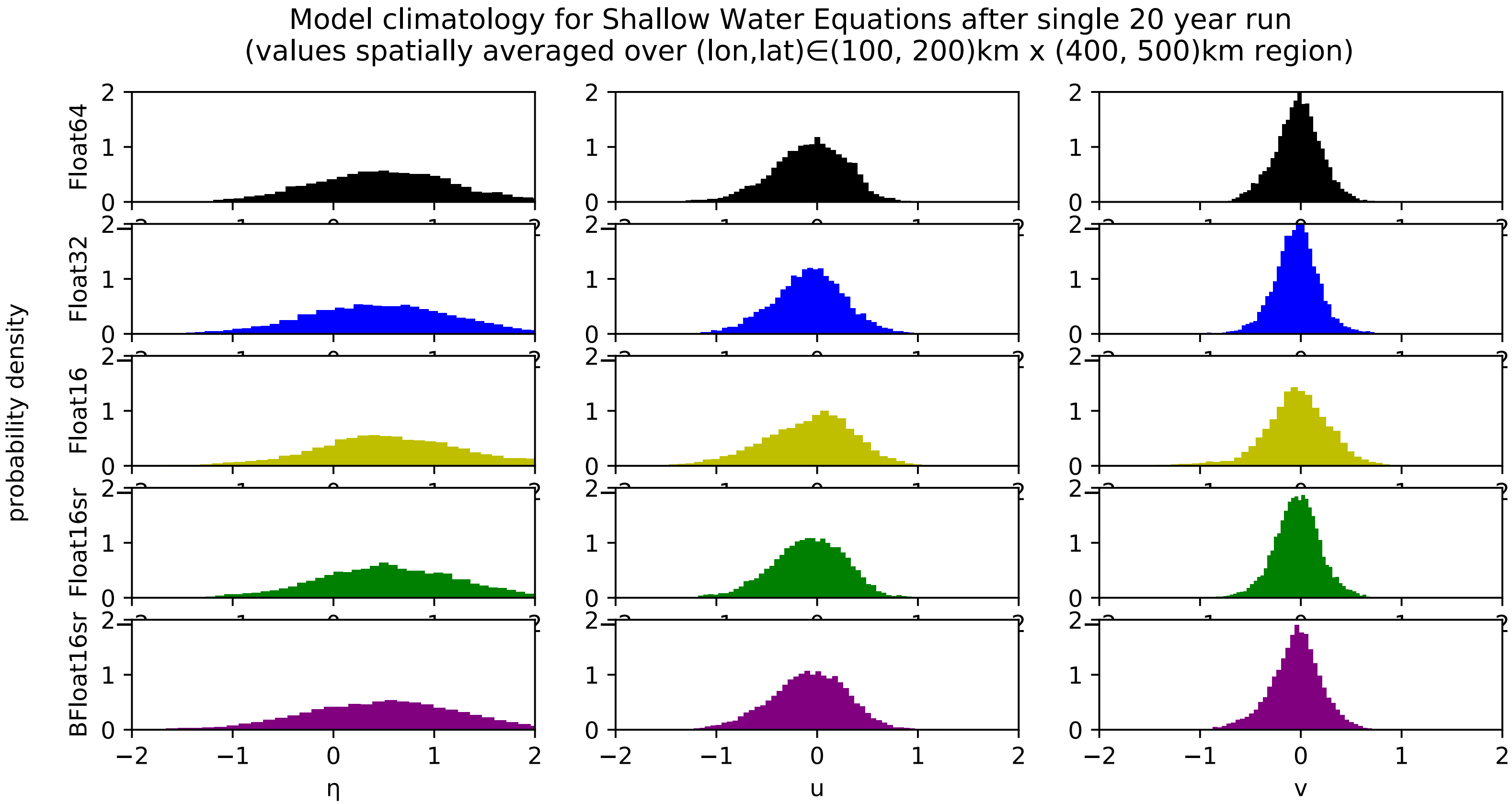
[4]Geosciences Department and Laboratoire de Météorologie Dynamique (CNRS and IPSL), Ecole Normale Supérieure and PSL University, Paris, France

[5]Department of Atmospheric & Oceanic Sciences, University of California at Los Angeles, Los Angeles, USA

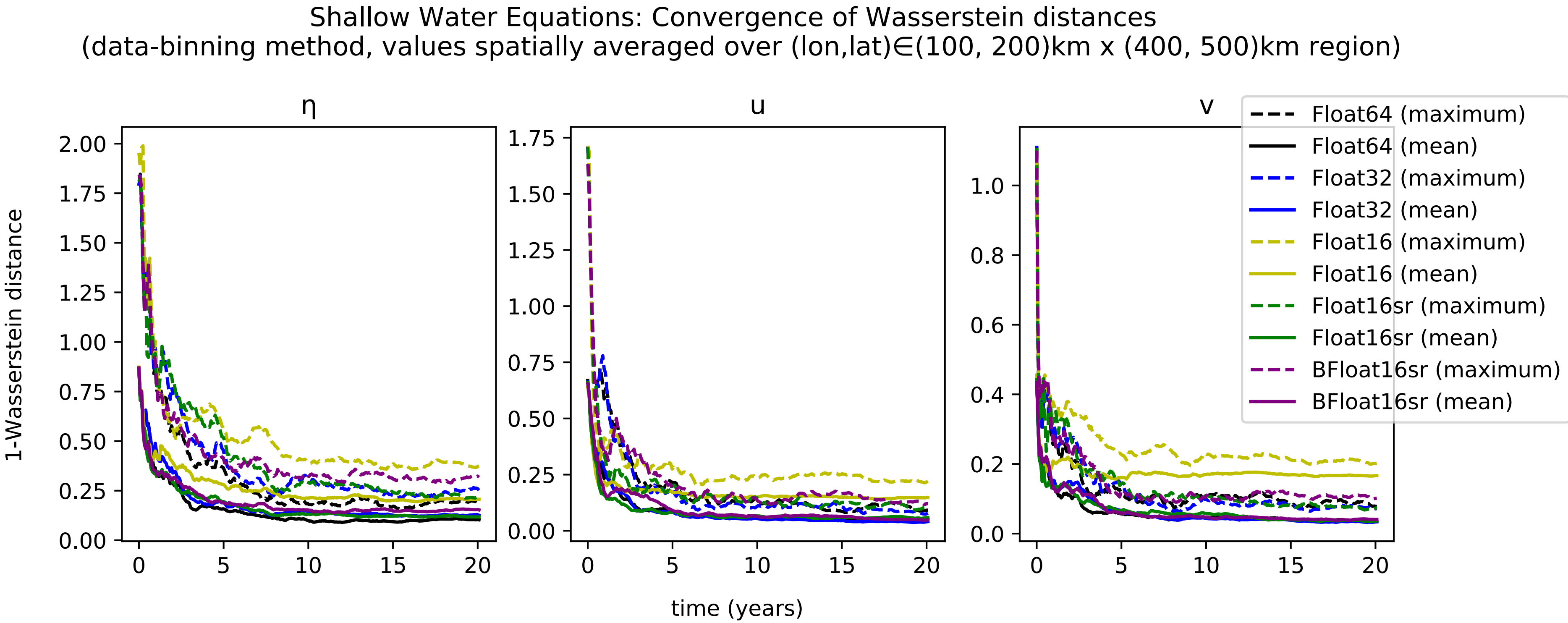
Key Points:

- Evaluation of climate model performance by benchmarking with reference datasets
- Climate model ranking related to the choice of variables of interest
- Highlighting model deficiencies through emphasis on climatic regions and variables

We can do this for Shallow Waters. Take spatial average over some (arbitrary) region (100,200)km x (400,500). Do 1D data-binning.



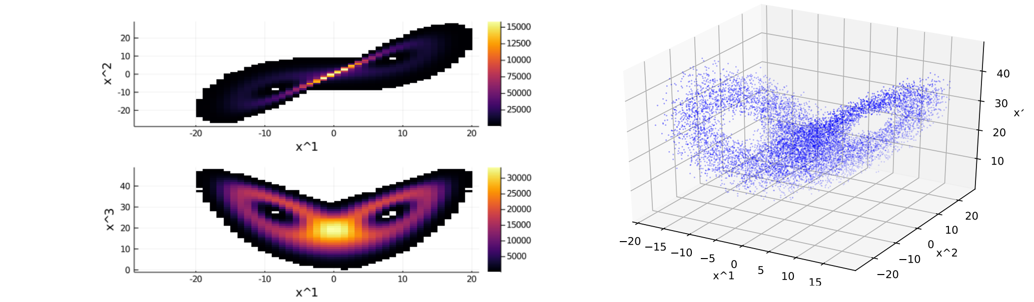
- We can compute Wasserstein distances between these 1D distributions.
- Same experiment as before (5-member ensembles, one Control ensemble).

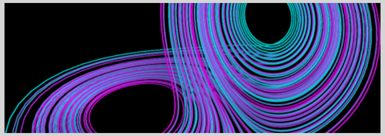


- The problem with projection is you are no longer considering the full distribution.

- The problem with projection is you are no longer considering the full distribution.
- IDEA: try the “scatter-plotting” method (direct sampling).

Recall: (a) data-binning, (b) scatter-plotting

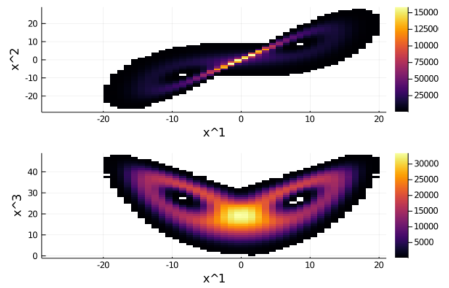


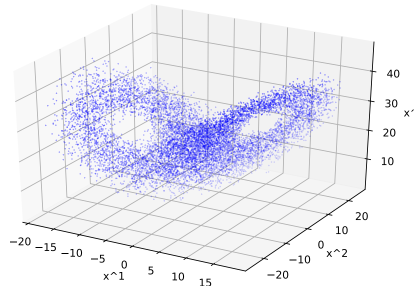


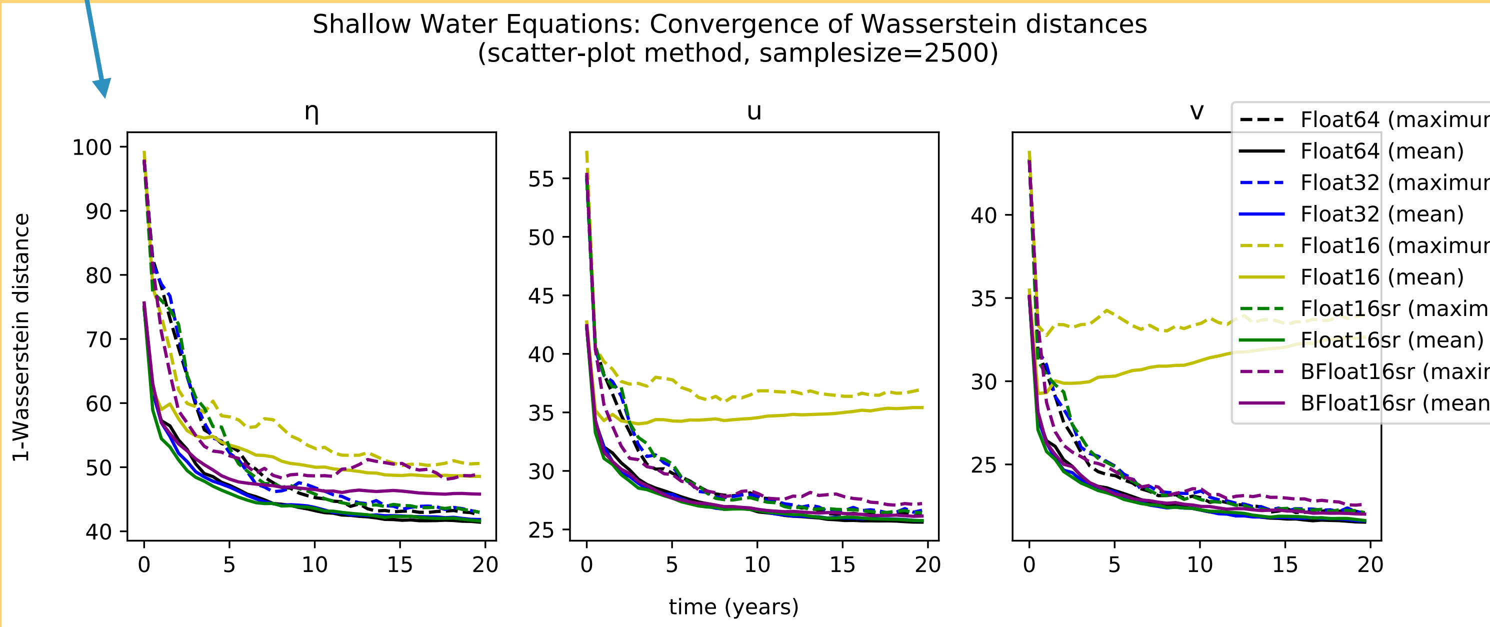
- The problem with projection is you are no longer considering the full distribution.
- IDEA: try the "scatter-plotting" method (direct sampling).

This seems to work!!!

Recall: (a) data-binning, (b) scatter-plotting

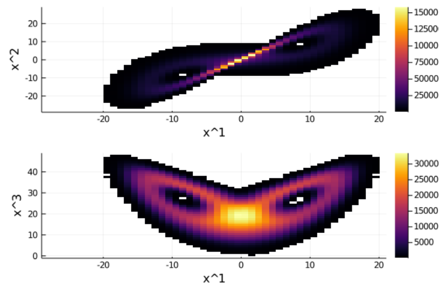


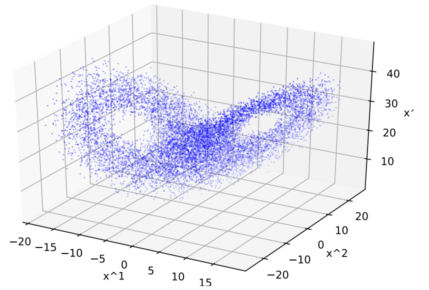




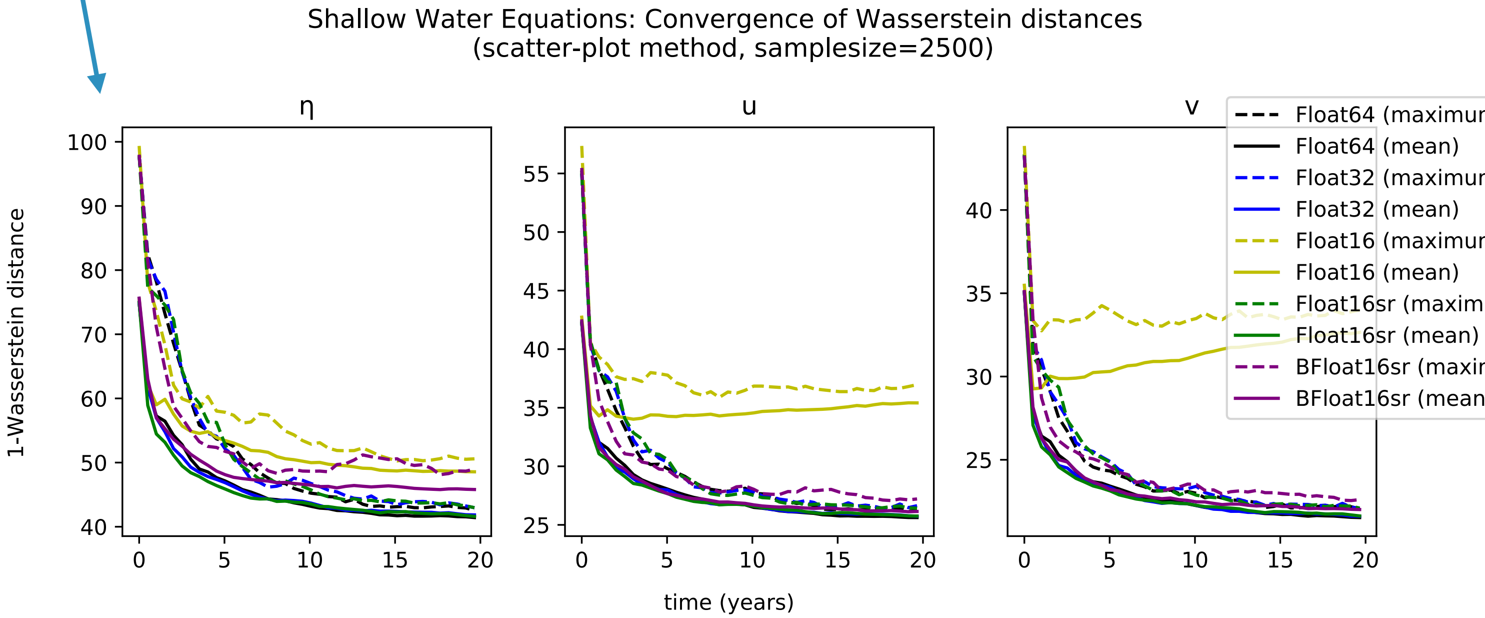
- The problem with projection is you are no longer considering the full distribution.
- IDEA: try the "scatter-plotting" method (direct sampling).

Recall: (a) data-binning, (b) scatter-plotting





This seems to work!!!

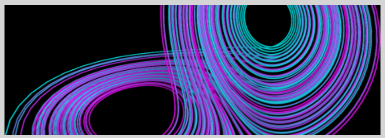


nb. This method is dimension agnostic (roughly the moral of Monte-Carlo methods)

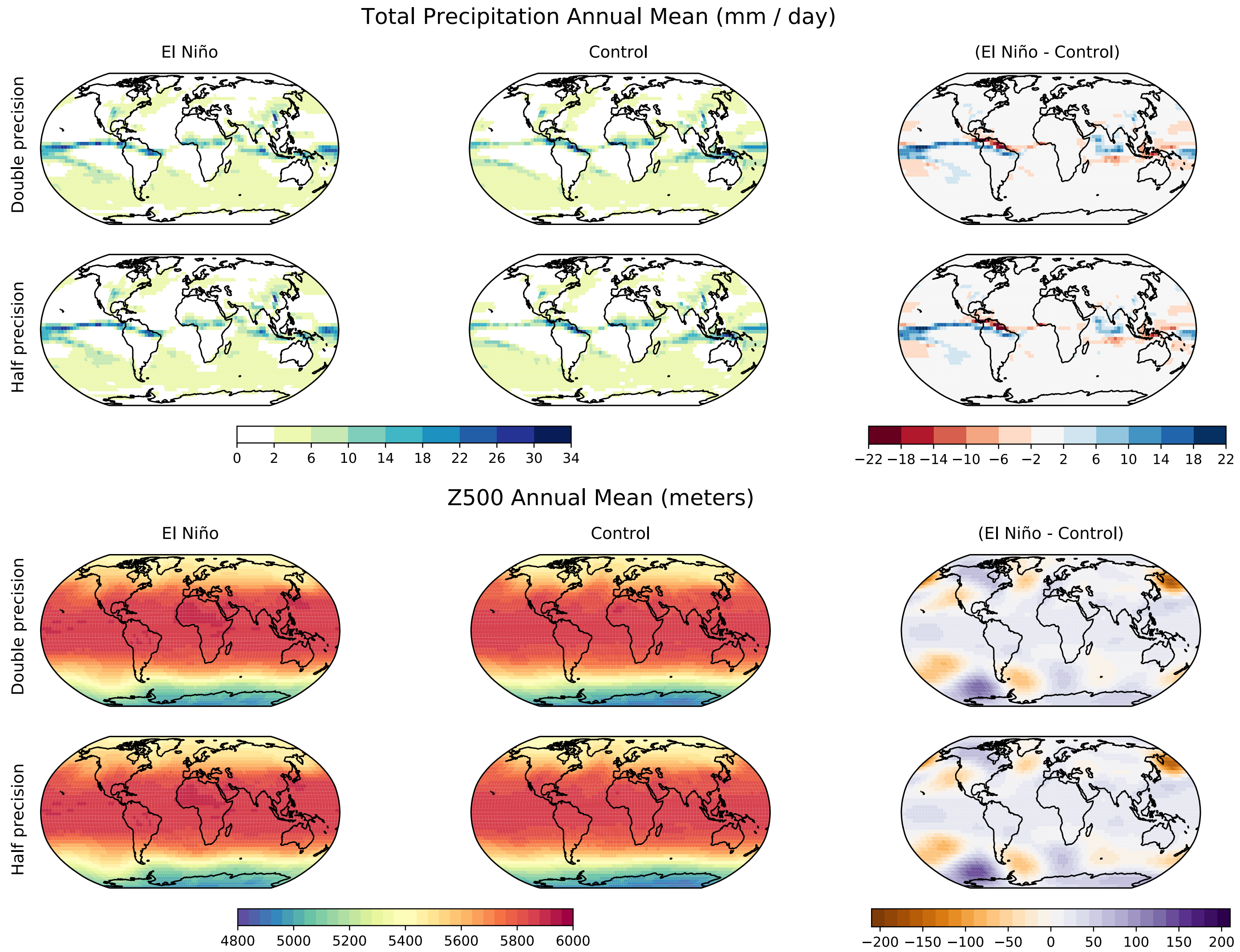
Conclusion of experiment.

The results provide strong evidence that the effects of rounding error on the shallow water model climatology, when compared with initial condition variability & discretisation error are:

1. *Negligible for **Float32** and **Float16sr**.*
2. *Significant for **Float16** and **BFloat16sr**.*

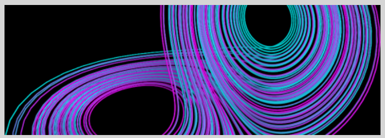


- Next steps: performing the same analysis to reduced precision SPEEDY.
- A coarse resolution ($3.75° \times 3.75°$) atmosphere only, primitive equation model (prescribed SSTs) with simplified parameterisations.
- Leo's 16-bit (deterministic) version of the code has held up to the first tests.



Summary of talk:

- The Wasserstein metric gives a notion of distance between probability distributions.
- It has excellent properties.
- It's computation presents challenges.
- Nonetheless it is a powerful tool for exploring high-dimensional probability distributions.
- Using the WD, the ensemble method, and ideas from sampling theory we have designed an experiment to test effects of rounding error on model climatology.
- Half-precision with stochastic-rounding is a suitable arithmetic for climate modelling with both of the L63 and Shallow Water models investigated so far.



Summary of talk:

- The Wasserstein metric gives a notion of distance between probability distributions.
- It has excellent properties.
- It's computation presents challenges.
- Nonetheless it is a powerful tool for exploring high-dimensional probability distributions.
- Using the WD, the ensemble method, and ideas from sampling theory we have designed an experiment to test effects of rounding error on model climatology.
- Half-precision with stochastic-rounding is a suitable arithmetic for climate modelling with both of the L63 and Shallow Water models investigated so far.

Thank-you!!! :)

... Any questions/thoughts/suggestions?