Look at the Data

Importance of Methods

The methods used to analyze time series are an essential factor in determining how much information can be extracted. But...

Importance of Methods

The methods used to analyze time series are an essential factor in determining how much information can be extracted. But...

- Not all methods yield useful results for any particular dataset.

Importance of Methods

The methods used to analyze time series are an essential factor in determining how much information can be extracted. But...

- Not all methods yield useful results for any particular dataset.

- It can take a lot of time and effort to learn a new method to the point of proficiency.

Importance of Methods

The methods used to analyze time series are an essential factor in determining how much information can be extracted. But...

- Not all methods yield useful results for any particular dataset.

- It can take a lot of time and effort to learn a new method to the point of proficiency.

- Creative use of low-tech methods can often be more fruitful than learning a new, supposedly sophisticated method.

Importance of Methods

The methods used to analyze time series are an essential factor in determining how much information can be extracted. But...

- Not all methods yield useful results for any particular dataset.

- It can take a lot of time and effort to learn a new method to the point of proficiency.

- Creative use of low-tech methods can often be more fruitful than learning a new, supposedly sophisticated method.

- The importance of ancillary data—that is, “adjacent” datasets that provide support and background— should not be overlooked.

Importance of Methods

The methods used to analyze time series are an essential factor in determining how much information can be extracted. But…

• Not all methods yield useful results for any particular dataset.

• It can take a lot of time and effort to learn a new method to the point of proficiency.

• Creative use of low-tech methods can often be more fruitful than learning a new, supposedly sophisticated method.

• The importance of ancillary data—that is, “adjacent” datasets that provide support and background— should not be overlooked.

• There is a limit to the amount of information that can be extracted from a given dataset. It is important not to take that personally.

Importance of Methods

The methods used to analyze time series are an essential factor in determining how much information can be extracted. But…

• Not all methods yield useful results for any particular dataset.

• It can take a lot of time and effort to learn a new method to the point of proficiency.

• Creative use of low-tech methods can often be more fruitful than learning a new, supposedly sophisticated method.

• The importance of ancillary data—that is, “adjacent” datasets that provide support and background— should not be overlooked.

• There is a limit to the amount of information that can be extracted from a given dataset. It is important not to take that personally.

• Our *attitudes* can matter as much as our *methods*.

Look at the Data

The first phase of analyzing a time series is to get to know it. This is done just by using your eyes.

Look at the Data

The first phase of analyzing a time series is to get to know it. This is done just by using your eyes.

***Looking at the data* is the single most important part of time series analysis.**

Look at the Data

The first phase of analyzing a time series is to get to know it. This is done just by using your eyes.

***Looking at the data* is the single most important part of time series analysis.**

Remarkably, many people skip over this step, and proceed straight to computing EOFs or spectra or whatever.

Look at the Data

The first phase of analyzing a time series is to get to know it. This is done just by using your eyes.

***Looking at the data* is the single most important part of time series analysis.**

Remarkably, many people skip over this step, and proceed straight to computing EOFs or spectra or whatever.

That's like starting to cook without even looking at your ingredients. It's unlikely to turn out well.

Look at the Data

The first phase of analyzing a time series is to get to know it. This is done just by using your eyes.

***Looking at the data* is the single most important part of time series analysis.**

Remarkably, many people skip over this step, and proceed straight to computing EOFs or spectra or whatever.

That's like starting to cook without even looking at your ingredients. It's unlikely to turn out well.

The goal of looking is to start to understand what kind of information might be, and might not be, present within the data.

Look at the Data

The first phase of analyzing a time series is to get to know it. This is done just by using your eyes.

***Looking at the data* is the single most important part of time series analysis.**

Remarkably, many people skip over this step, and proceed straight to computing EOFs or spectra or whatever.

That's like starting to cook without even looking at your ingredients. It's unlikely to turn out well.

The goal of looking is to start to understand what kind of information might be, and might not be, present within the data.

In other words, you'd like to know what the data is “trying” to say.

Some Guidelines

Analyzing data is about being an *observationalist*. This means freeing your mind from preconceptions and being able to notice what is actually present.

Some Guidelines

Analyzing data is about being an *observationalist*. This means freeing your mind from preconceptions and being able to notice what is actually present.

The observational scientist is the person who *speaks for* the data. This an an *impersonal* activity, unrelated to your wishes and goals.

Some Guidelines

Analyzing data is about being an *observationalist*. This means freeing your mind from preconceptions and being able to notice what is actually present.

The observational scientist is the person who *speaks for* the data. This an an *impersonal* activity, unrelated to your wishes and goals.

Use your imagination. You are putting together a puzzle with only 10% of the pieces. What does the whole puzzle look like? The data can *support* a hypothesis much larger than the data themselves.

Some Guidelines

Analyzing data is about being an *observationalist*. This means freeing your mind from preconceptions and being able to notice what is actually present.

The observational scientist is the person who *speaks for* the data. This an an *impersonal* activity, unrelated to your wishes and goals.

Use your imagination. You are putting together a puzzle with only 10% of the pieces. What does the whole puzzle look like? The data can *support* a hypothesis much larger than the data themselves.

It is important to assess limitations realistically. Be honest! The observational scientist's job includes being able to recongize when something is not clear or not supported.

Looking at the Data

The first thing we will do is to practice looking at data.

Looking at the Data

The first thing we will do is to practice looking at data.

Try to note as many *observable features* as you can before moving on to the next slide.

Looking at the Data

The first thing we will do is to practice looking at data.

Try to note as many *observable features* as you can before moving on to the next slide.

Ask yourself what you can know, and not know, just by seeing.

Looking at the Data

The first thing we will do is to practice looking at data.

Try to note as many *observable features* as you can before moving on to the next slide.

Ask yourself what you can know, and not know, just by seeing.

You are not told what the data is.

Looking at the Data

The first thing we will do is to practice looking at data.

Try to note as many *observable features* as you can before moving on to the next slide.

Ask yourself what you can know, and not know, just by seeing.

You are not told what the data is.

You are also not provided with important information, like units.

Looking at the Data

The first thing we will do is to practice looking at data.

Try to note as many *observable features* as you can before moving on to the next slide.

Ask yourself what you can know, and not know, just by seeing.

You are not told what the data is.

You are also not provided with important information, like units.

A grey box denotes a region that will be zoomed in on the next slide.

Looking at the Data

The first thing we will do is to practice looking at data.

Try to note as many *observable features* as you can before moving on to the next slide.

Ask yourself what you can know, and not know, just by seeing.

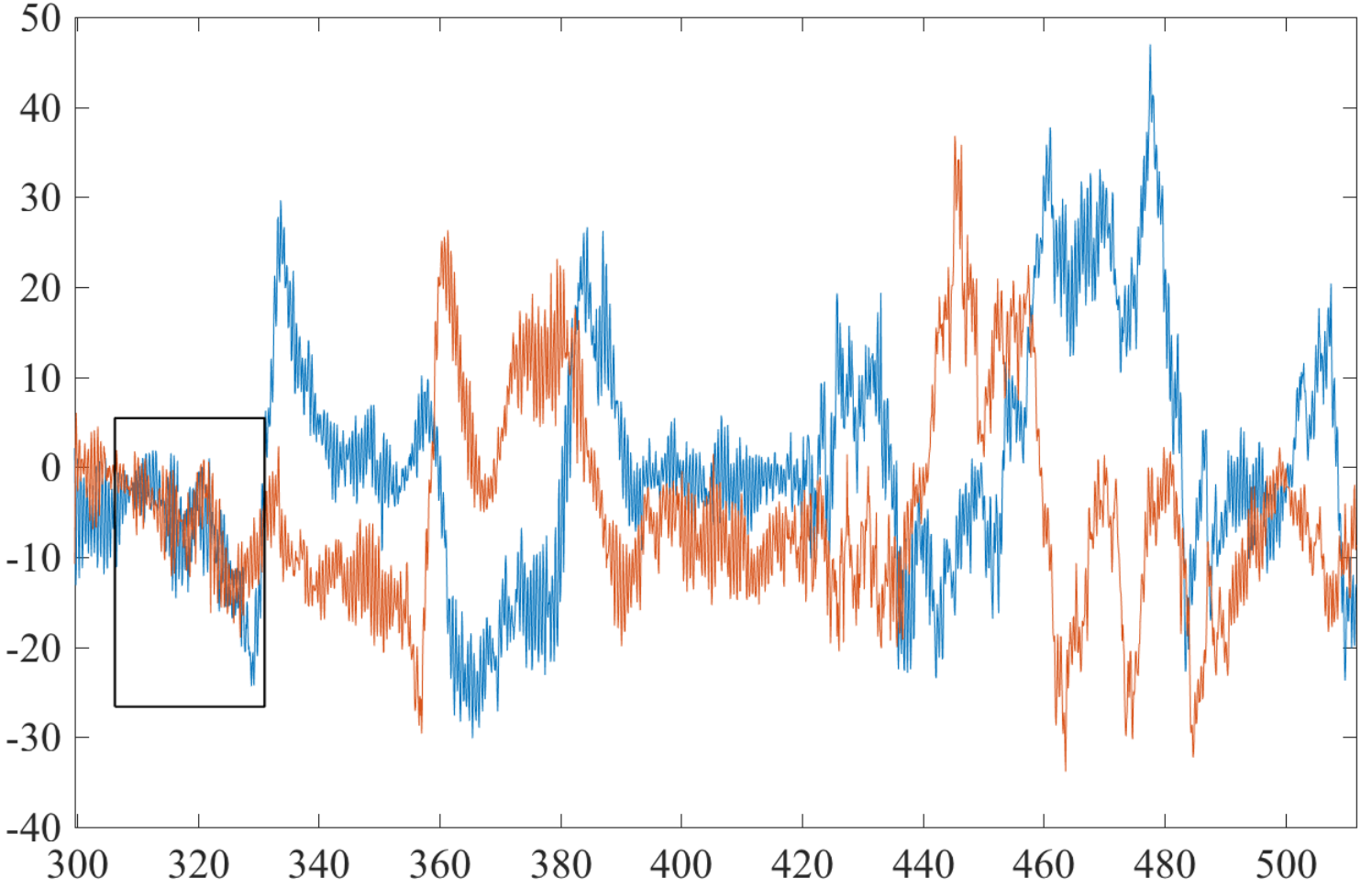
You are not told what the data is.

You are also not provided with important information, like units.

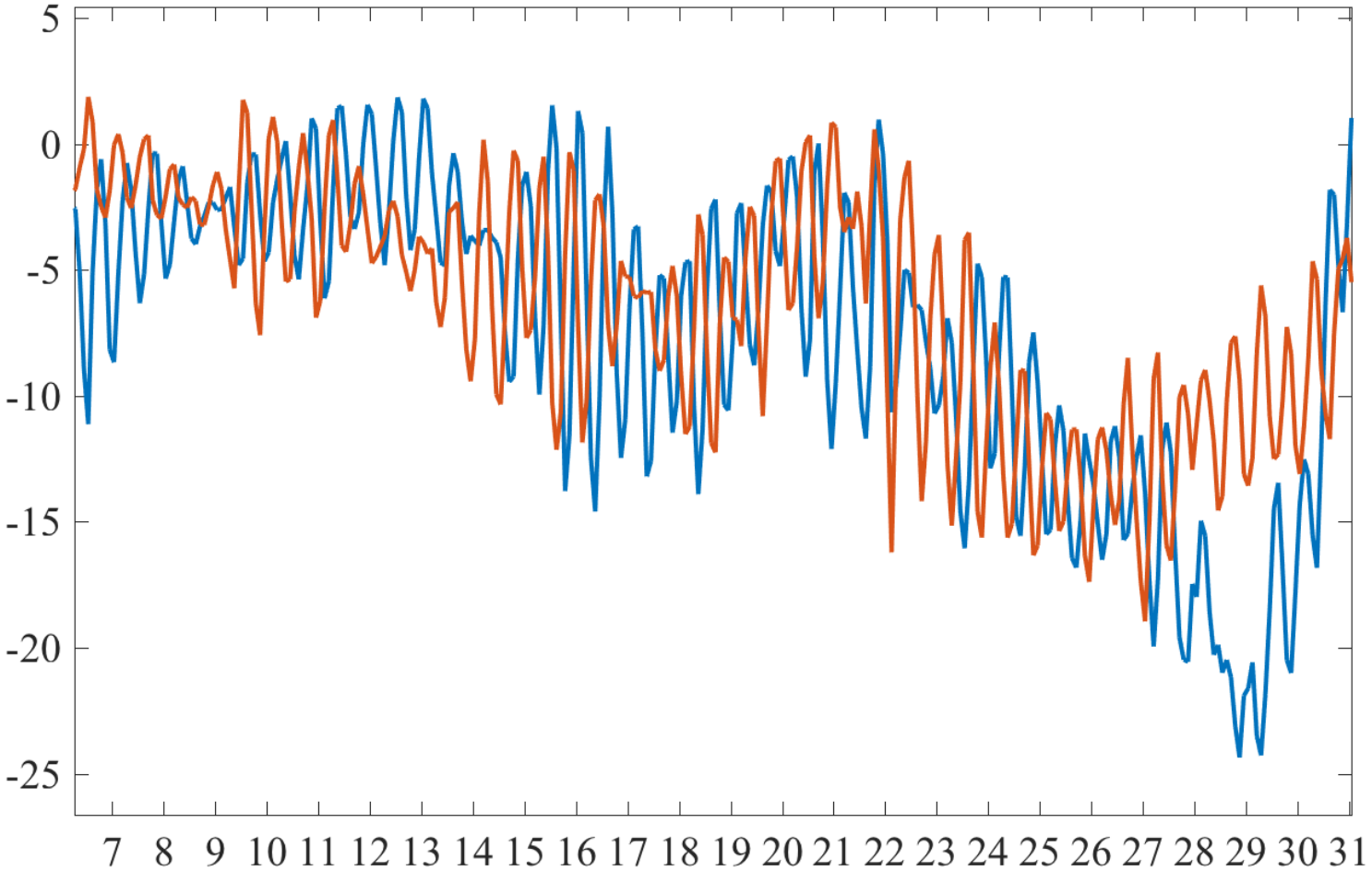
A grey box denotes a region that will be zoomed in on the next slide.

After you have noted as many features as you can, see if you can guess what the data might be.

First Example



First Example



Observable Features

1. The data consists of two time series that are similar in character.
2. Both time series present a superposition of scales.
3. At the smallest scale, there is an apparently oscillatory roughness which changes its amplitude in time.
4. A larger scale presents itself either as localized features, or as wavelike in nature.
5. Several sudden transitions are associated with isolated events.
6. Zooming in, we see the small-scale oscillatory behavior is sometimes 90° degrees out of phase, and sometimes 180°.
7. The amplitude of this oscillatory variability changes with time.

The fact that the oscillatory behavior is not consistently 90° out of phase removes the possibility of these features being purely inertial oscillations. The amplitude modulation suggests tidal beating.

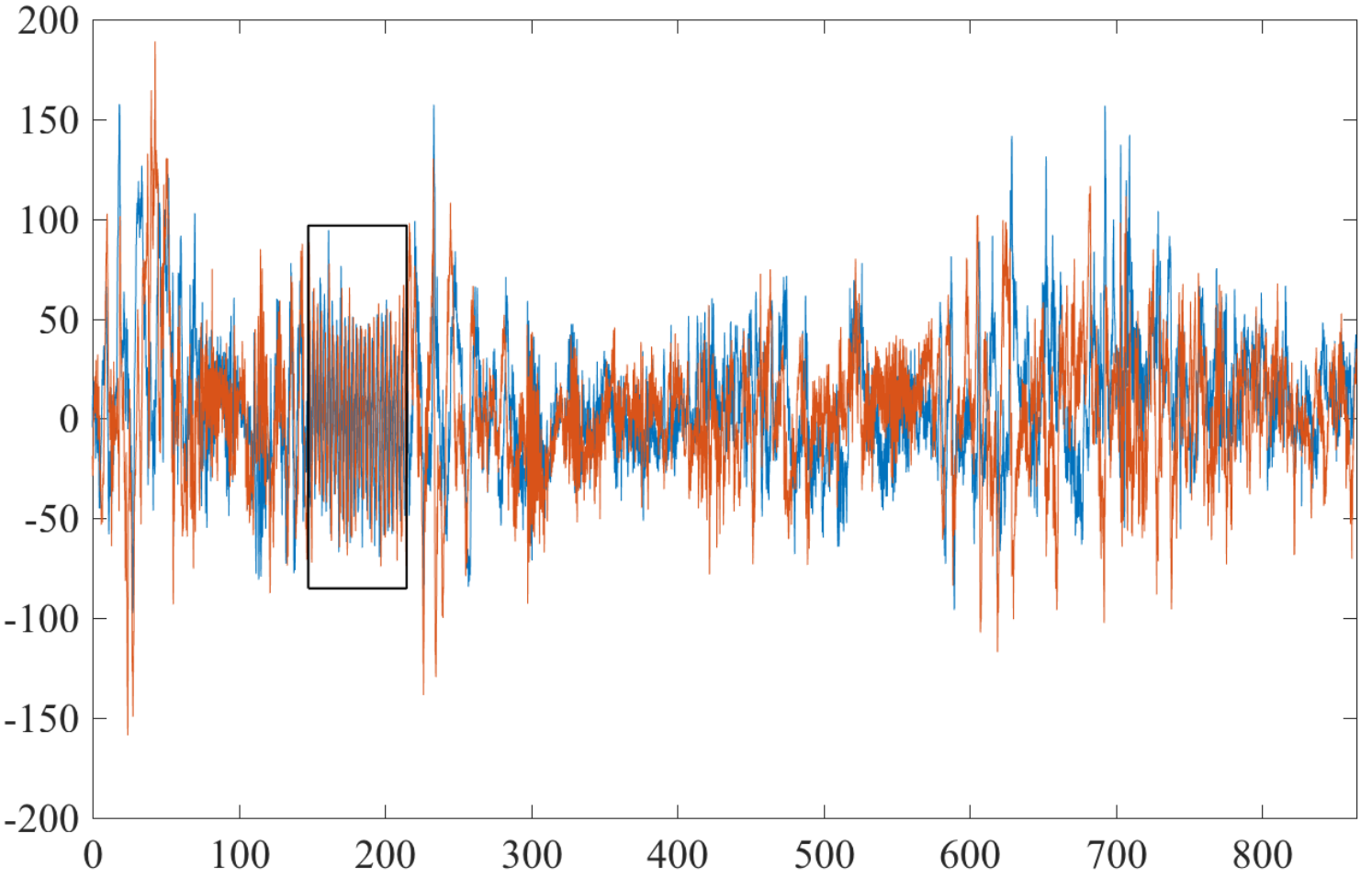
Observable Features

1. The data consists of two time series that are similar in character.
2. Both time series present a superposition of scales.
3. At the smallest scale, there is an apparently oscillatory roughness which changes its amplitude in time.
4. A larger scale presents itself either as localized features, or as wavelike in nature.
5. Several sudden transitions are associated with isolated events.
6. Zooming in, we see the small-scale oscillatory behavior is sometimes 90° degrees out of phase, and sometimes 180°.
7. The amplitude of this oscillatory variability changes with time.

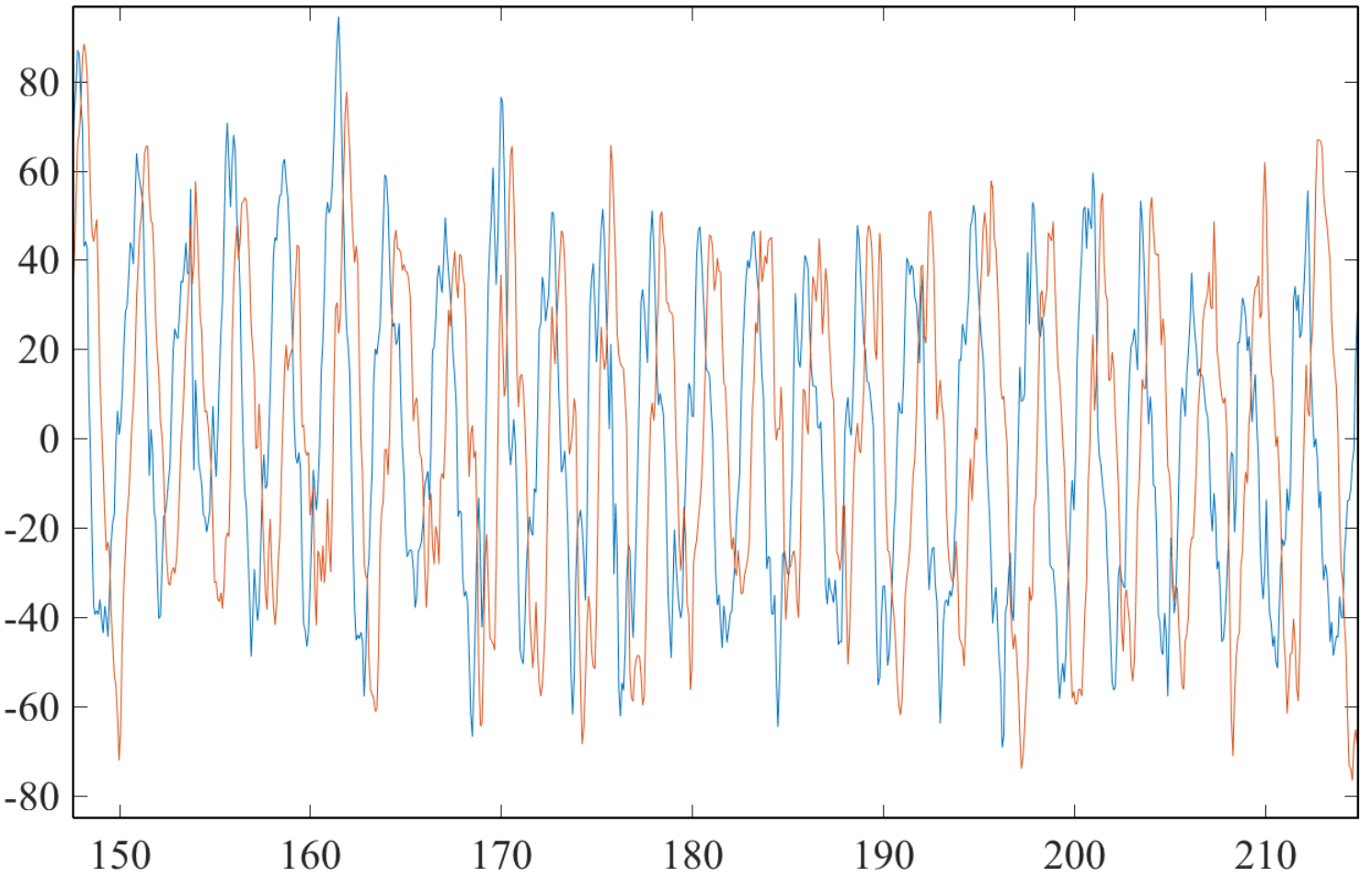
The fact that the oscillatory behavior is not consistently 90° out of phase removes the possibility of these features being purely inertial oscillations. The amplitude modulation suggests tidal beating.

This is current meter data from the Labrador Sea. The isolated events are eddies, which cause the currents to suddenly rotate as they pass by. The oscillations are due to tides and internal waves.

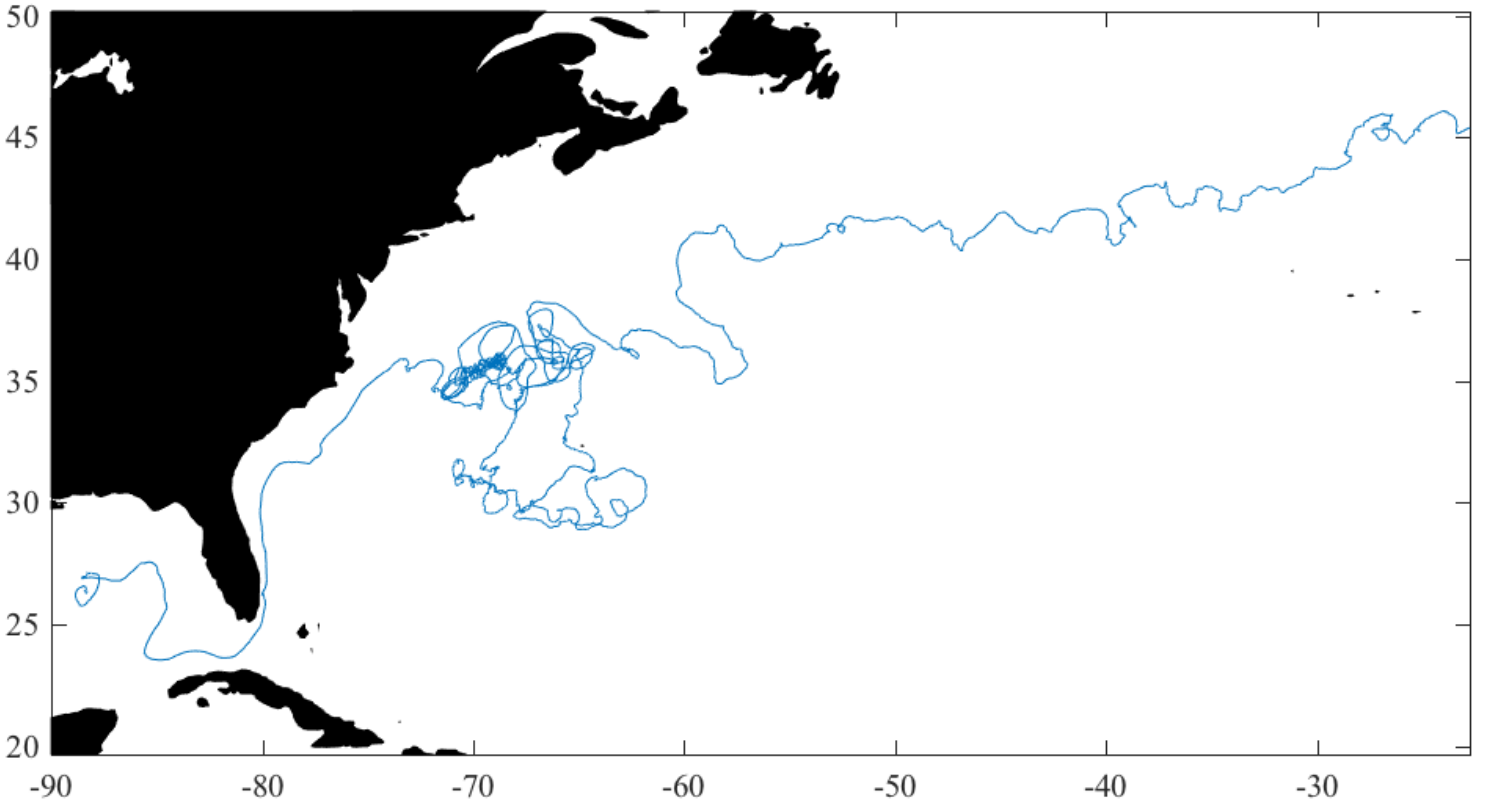
Second Example



Second Example



Second Example



Observable Features

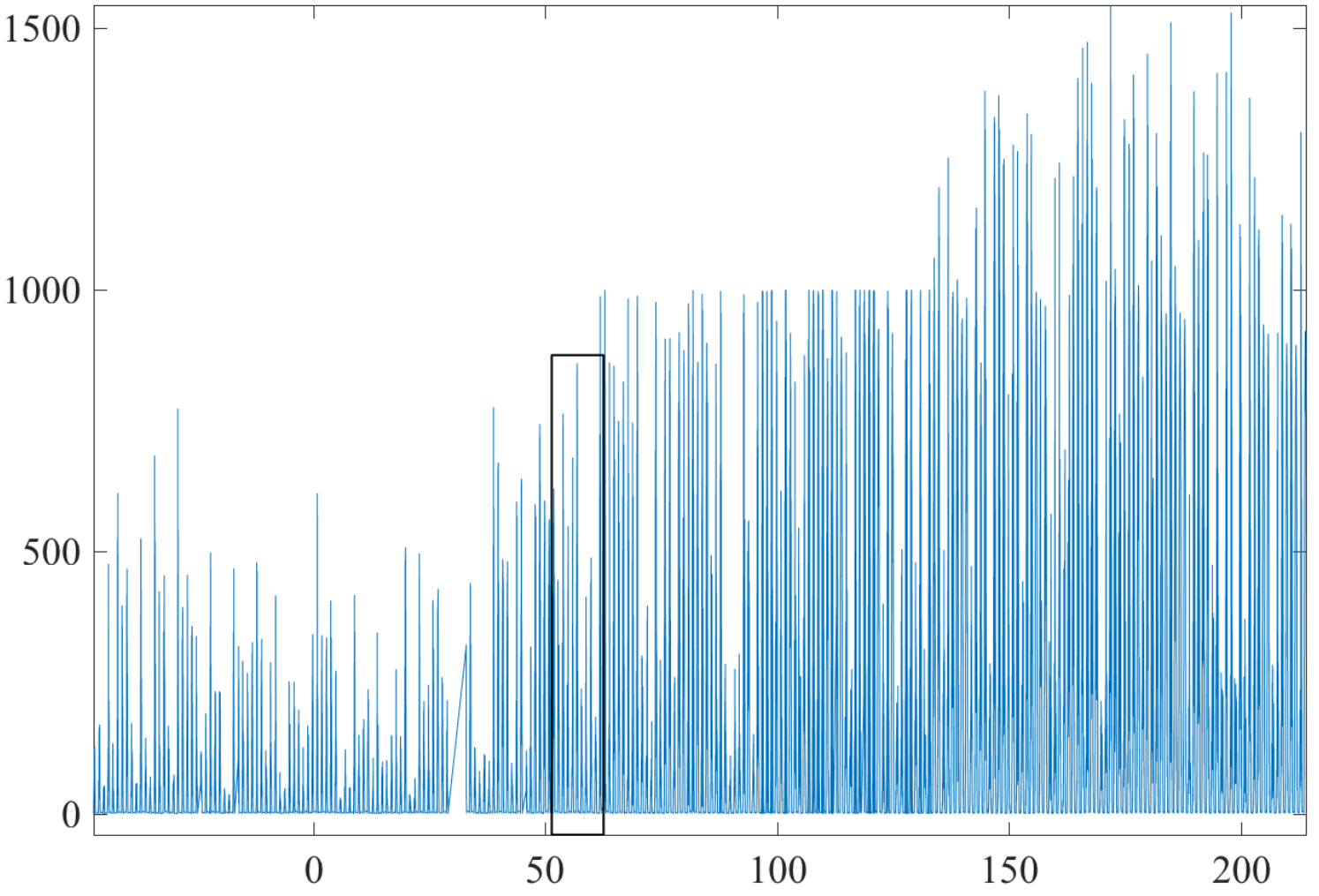
1. The data consists of two time series that are similar in character.
2. Both time series present a superposition of scales and a high degree of roughness.
3. The data seems to consist of different time periods with distinct statistical characteristics—the data is *nonstationary*.
4. Zooming in to one particular period show regular oscillations of roughly uniform amplitude and frequency.
5. The phasing of these show a circular polarization orbited in a counterclockwise direction.
6. The zoomed-in plot shows a fair amount of what appears to be measurement noise superimposed on the oscillatory signal.

Observable Features

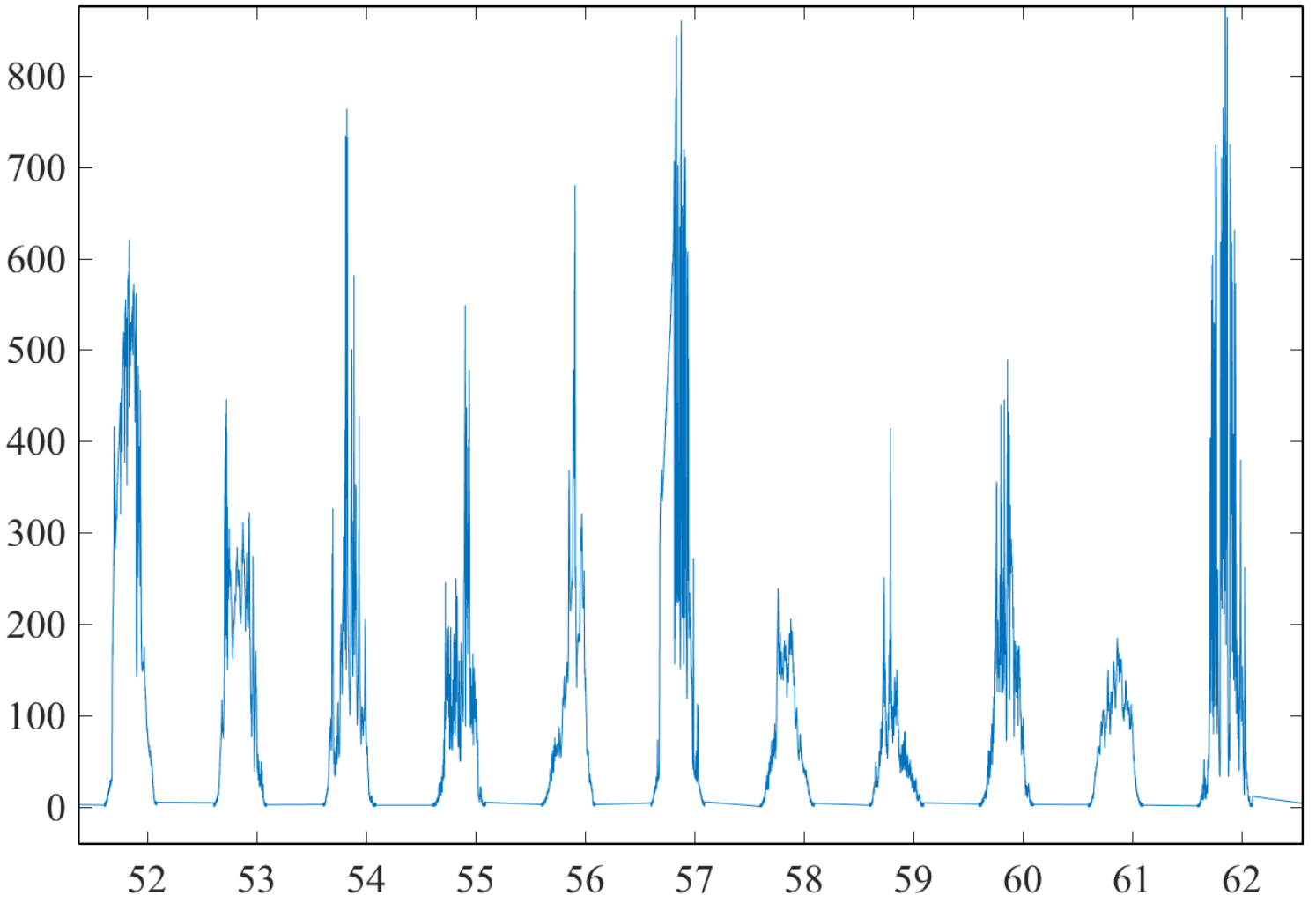
1. The data consists of two time series that are similar in character.
2. Both time series present a superposition of scales and a high degree of roughness.
3. The data seems to consist of different time periods with distinct statistical characteristics—the data is *nonstationary*.
4. Zooming in to one particular period show regular oscillations of roughly uniform amplitude and frequency.
5. The phasing of these show a circular polarization orbited in a counterclockwise direction.
6. The zoomed-in plot shows a fair amount of what appears to be measurement noise superimposed on the oscillatory signal.

This is a surface drifter record. The oscillatory portion is due to trapping in a cyclonic eddy.

Third Example



Third Example



Observable Features

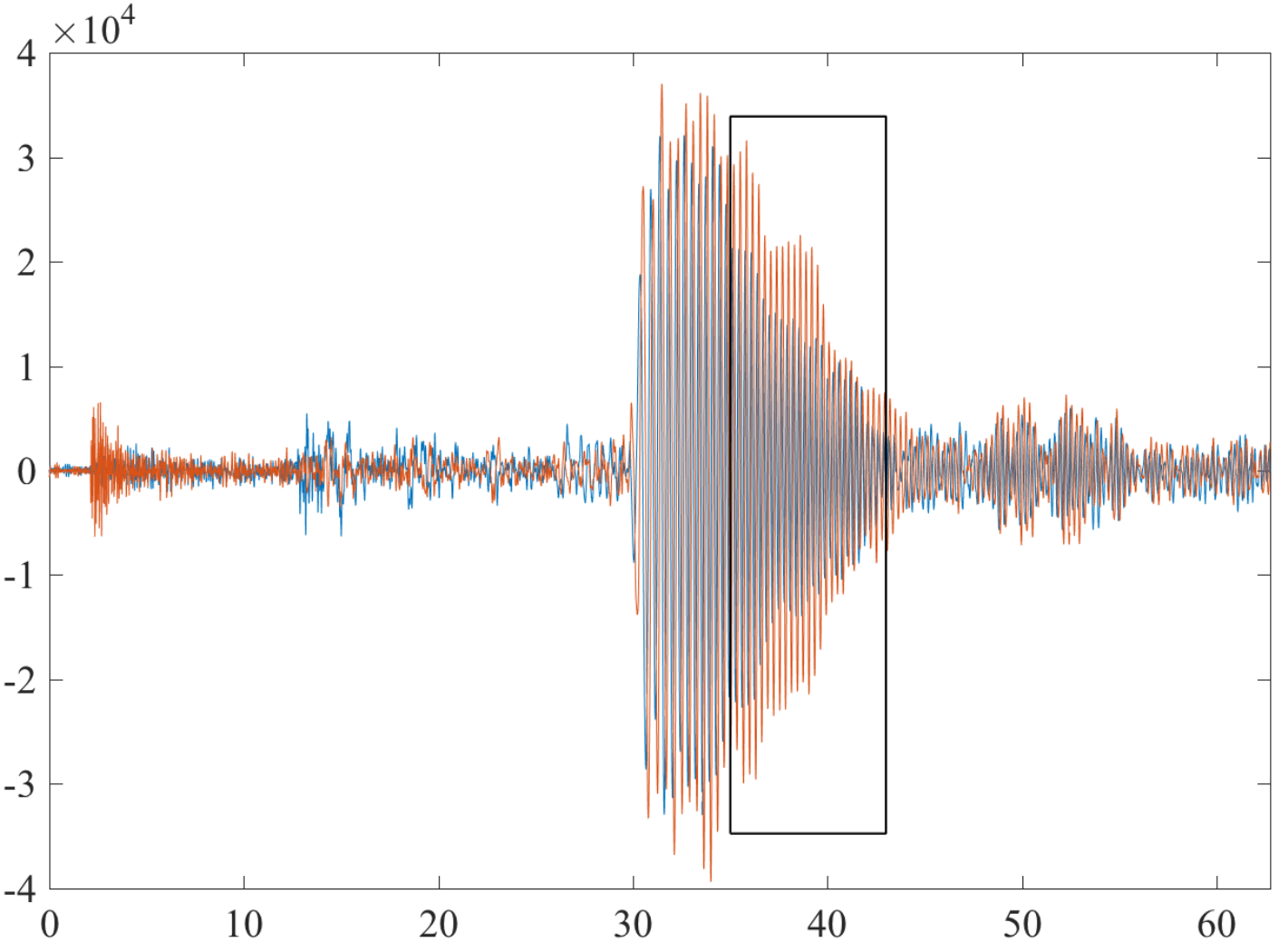
1. The data appears to be composed of nonnegative spikes at regularly spaced intervals.
2. The amplitude of the spikes generally increases over time.
3. During the middle part of the record, the amplitude conspicuously appears to obtain a fixed maximum value.
4. A time period of linearly increasing values is apparent, suggesting a gap filled by interpolation.
5. Zooming in shows that the data is composed of alternating periods of roughly *zero* values, and periods of positive values.
6. These two periods are of roughly equal length.
7. The positive-value periods are roughly symmetric, increasing to a maximum value near their midpoint before decaying again.
8. High-frequency variability is seen within the positive regions.

Observable Features

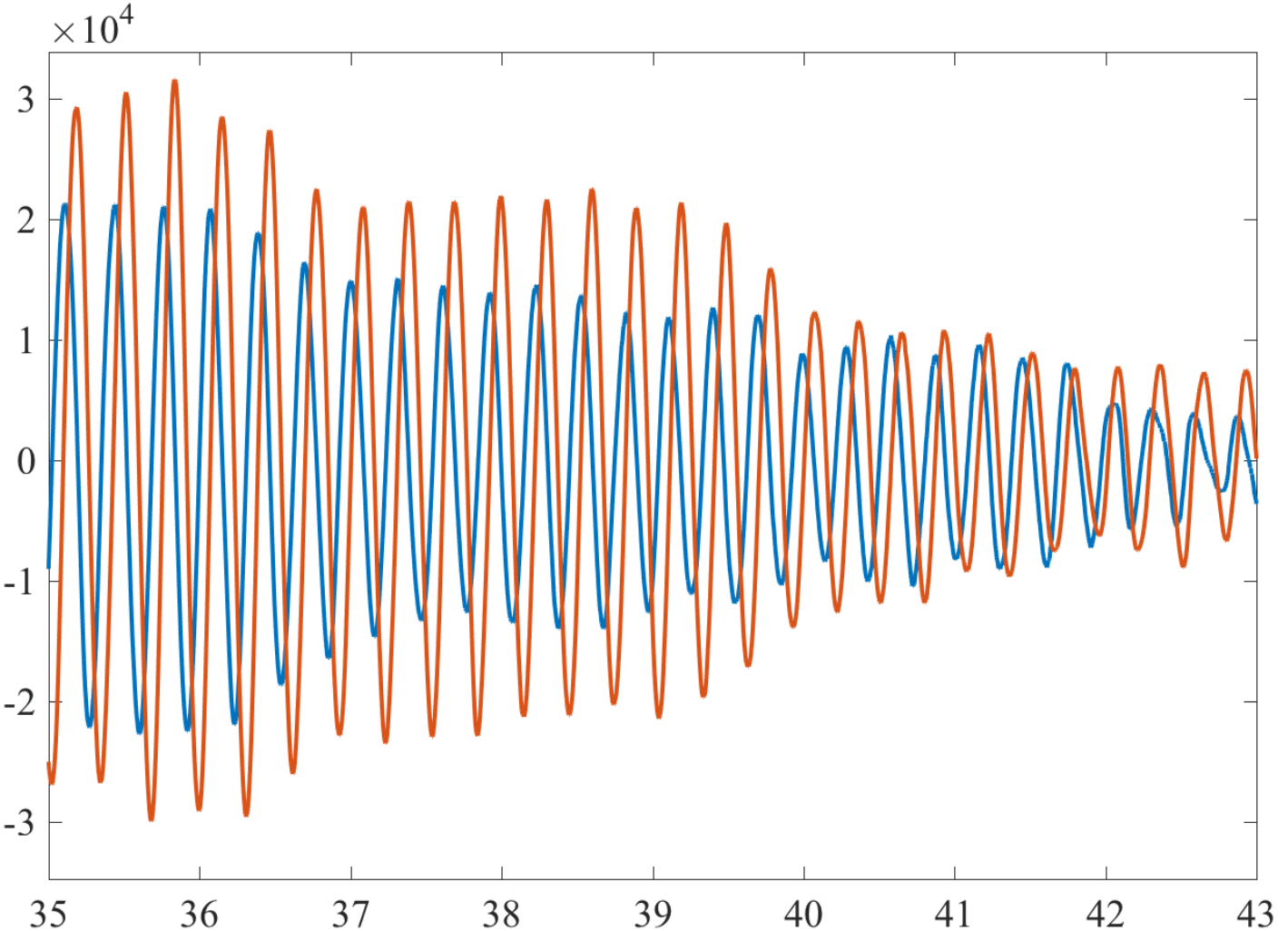
1. The data appears to be composed of nonnegative spikes at regularly spaced intervals.
2. The amplitude of the spikes generally increases over time.
3. During the middle part of the record, the amplitude conspicuously appears to obtain a fixed maximum value.
4. A time period of linearly increasing values is apparent, suggesting a gap filled by interpolation.
5. Zooming in shows that the data is composed of alternating periods of roughly *zero* values, and periods of positive values.
6. These two periods are of roughly equal length.
7. The positive-value periods are roughly symmetric, increasing to a maximum value near their midpoint before decaying again.
8. High-frequency variability is seen within the positive regions.

This is solar radiation data recorded from the roof of NWRA. We are seeing the diurnal cycle as well as the annual cycle. The uniform upper bound in the middle portion is suspicious and likely a data quality issue. The high-frequency variability is from passing clouds.

Fourth Example



Fourth Example



Observable Features

1. A very small intrinsic noise level, as seen at the beginning.
2. A sudden wave arrival near the beginning of the record.
3. A much larger wave arrival in the middle of the record.
4. From the phasing of the large wave arrival, we can see that it is *elliptically polarized*, with an eccentricity that changes in time. The orbital motion is in the *counterclockwise* sense.
5. There is no sign of asymmetry in the orbital motion, neither peak-to-trough nor left-to-right.
6. The characters of the early and late waves are very different. The early wave appears *jagged* while the later wave appears smooth.

Thus we appear to be seeing some kind of wave arrival, though the medium does not appear to be water. We need a medium that can support different types of waves.

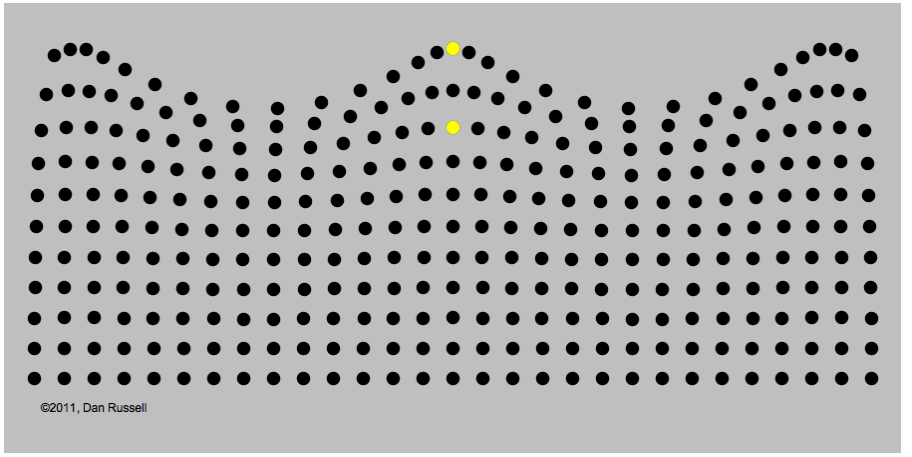
Observable Features

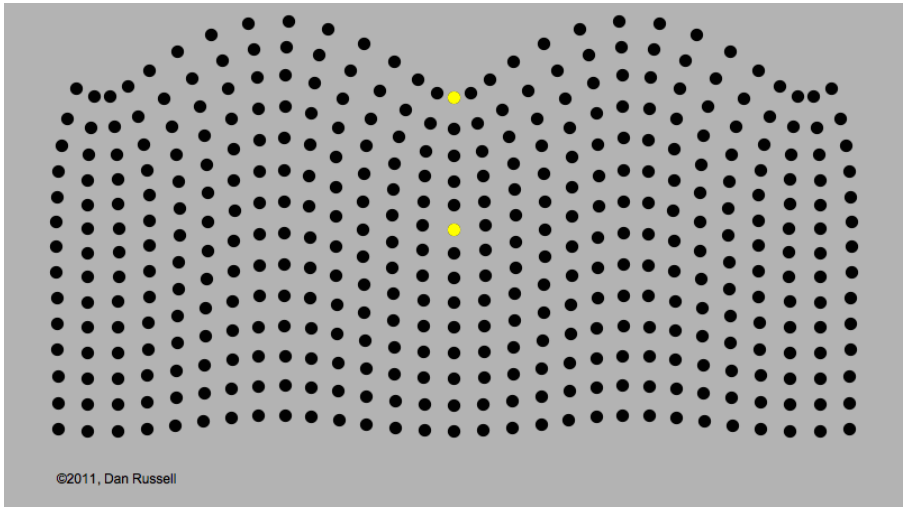
1. A very small intrinsic noise level, as seen at the beginning.
2. A sudden wave arrival near the beginning of the record.
3. A much larger wave arrival in the middle of the record.
4. From the phasing of the large wave arrival, we can see that it is *elliptically polarized*, with an eccentricity that changes in time. The orbital motion is in the *counterclockwise* sense.
5. There is no sign of asymmetry in the orbital motion, neither peak-to-trough nor left-to-right.
6. The characters of the early and late waves are very different. The early wave appears *jagged* while the later wave appears smooth.

Thus we appear to be seeing some kind of wave arrival, though the medium does not appear to be water. We need a medium that can support different types of waves.

This is a seismograph. We are looking at the radial (away from source in the horizontal plane) and vertical components of acceleration. The major wave is called a Rayleigh wave.

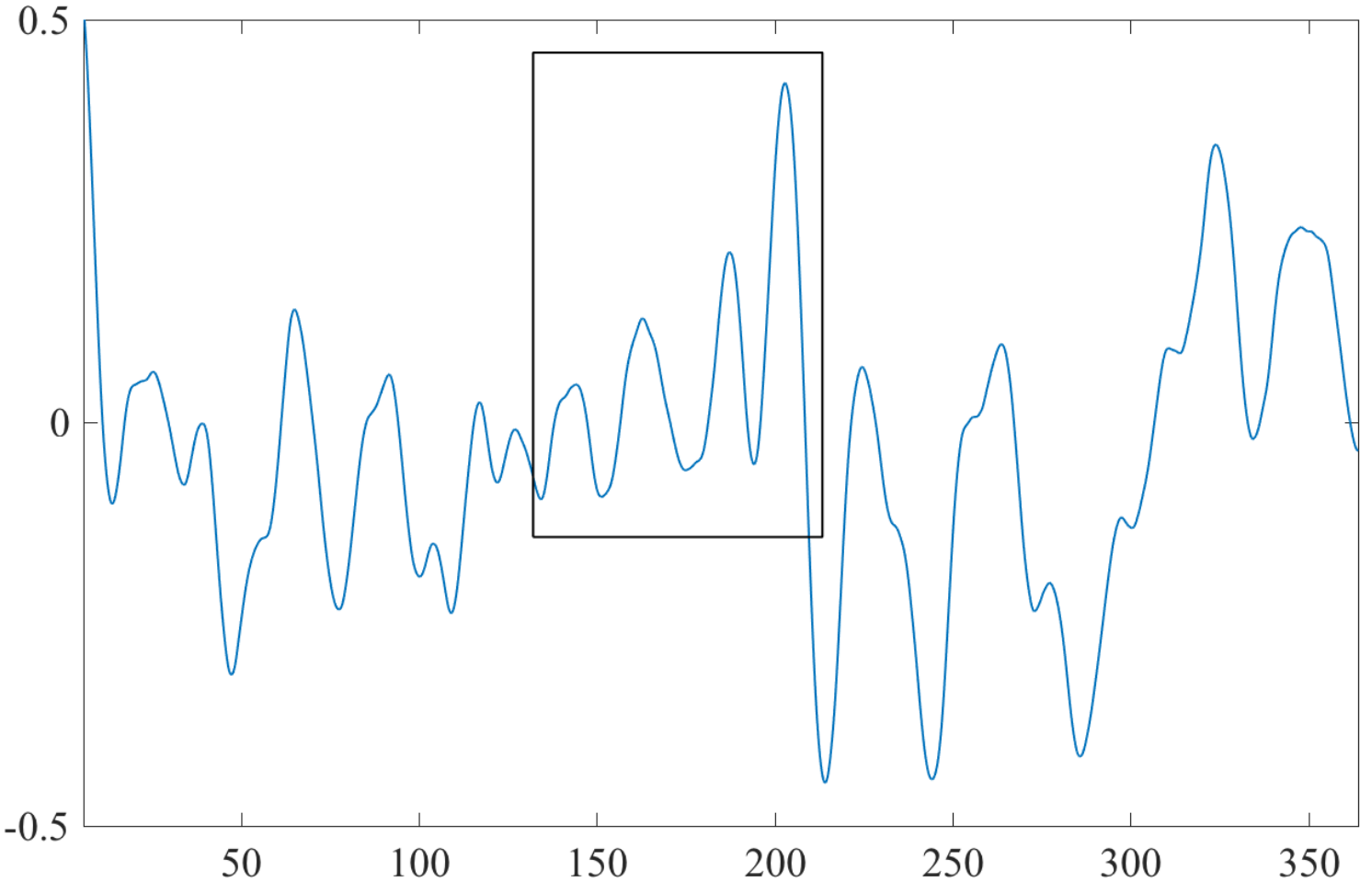
Water vs. Rayleigh Wave



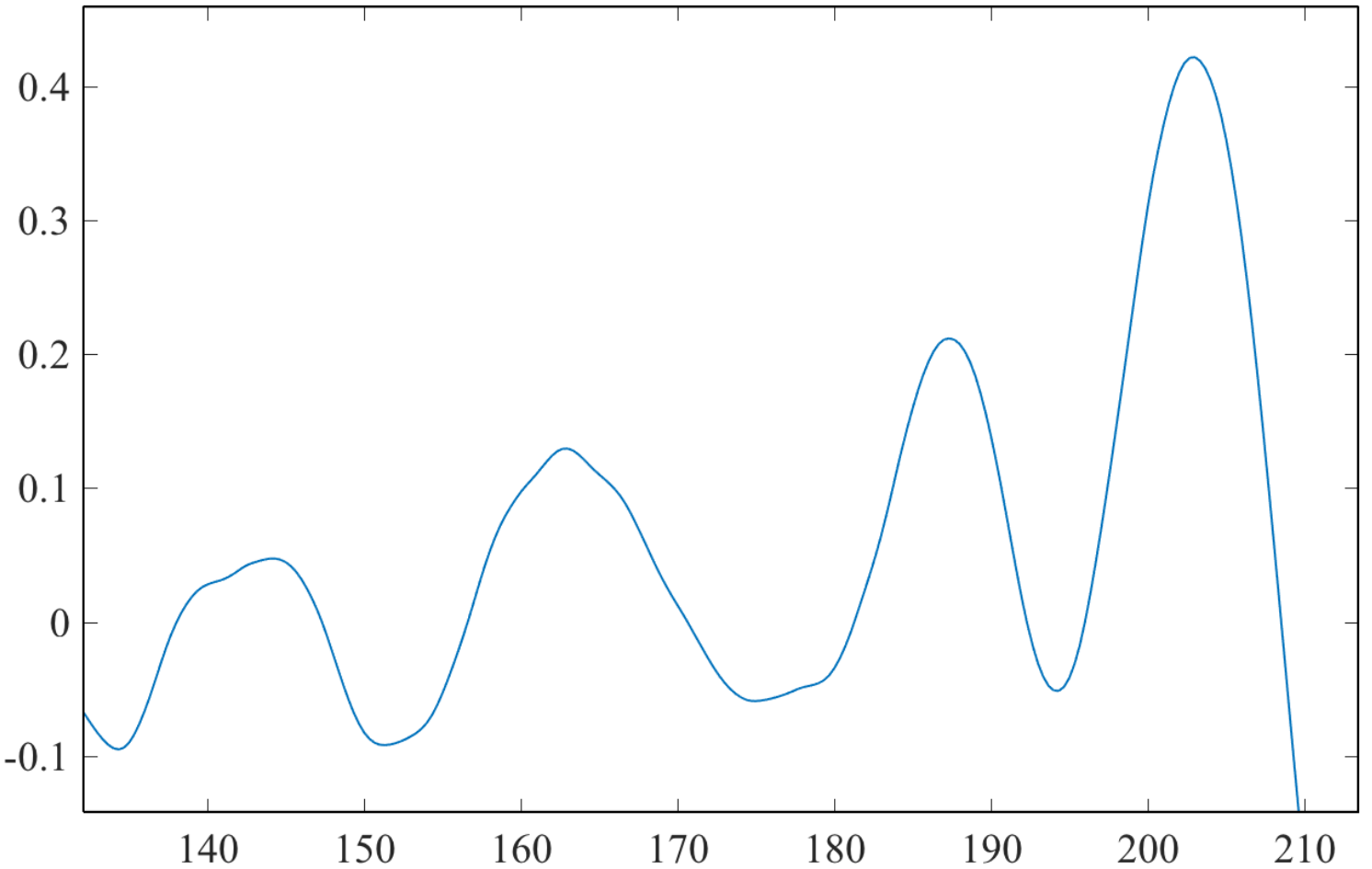


Thanks to {Dan Russell}.

Fifth Example



Fifth Example



Observable Features

1. This time series is smooth, suggesting it has been previously filtered.
2. The predominant variability is present is at a roughly 25—100 day time scale.
3. The amplitude of this variability is more or less uniform over the time interval.
4. An event near the center of the record appears to increase its amplitude as its frequency decreases.

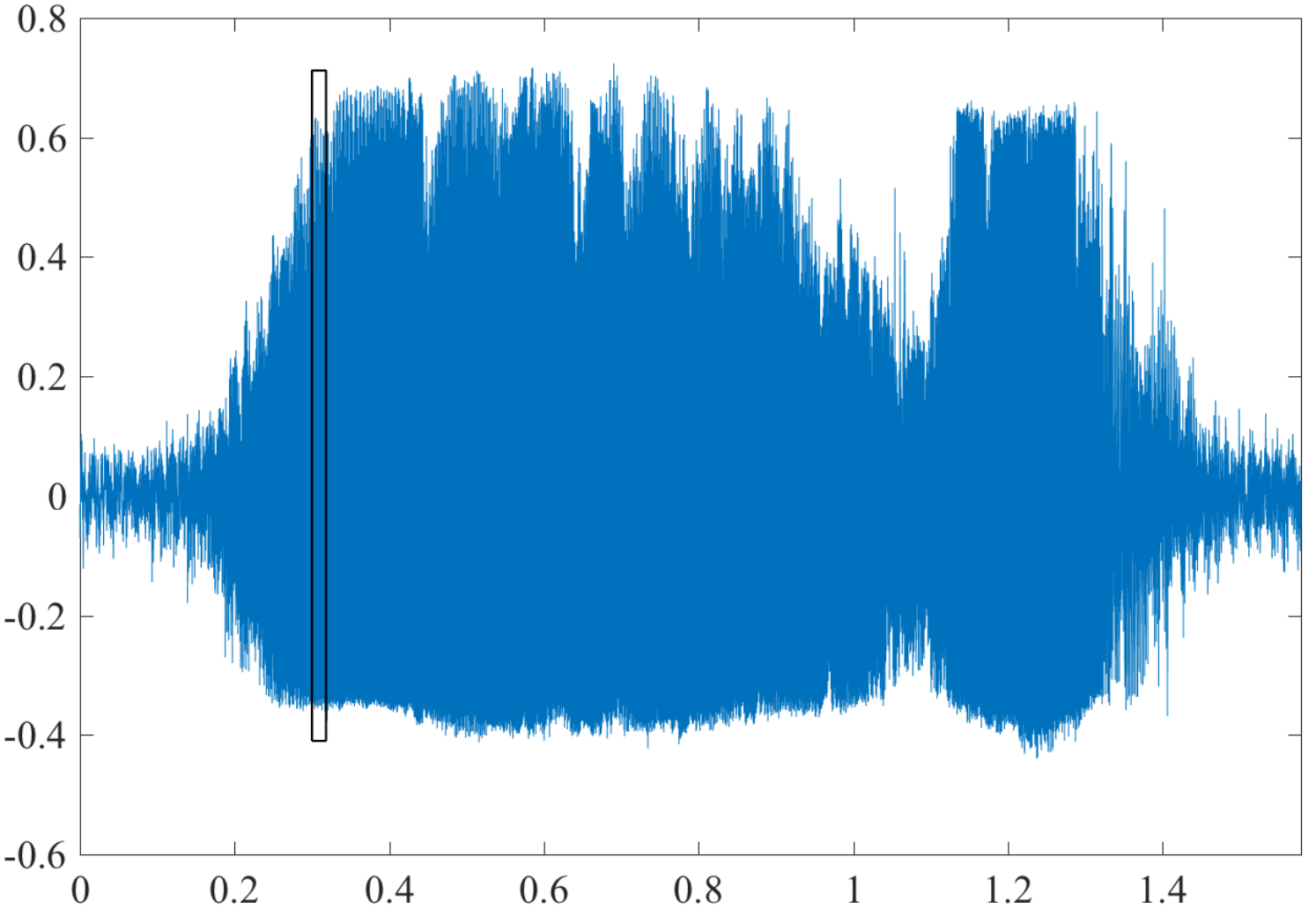
Observable Features

1. This time series is smooth, suggesting it has been previously filtered.
2. The predominant variability is present is at a roughly 25—100 day time scale.
3. The amplitude of this variability is more or less uniform over the time interval.
4. An event near the center of the record appears to increase its amplitude as its frequency decreases.

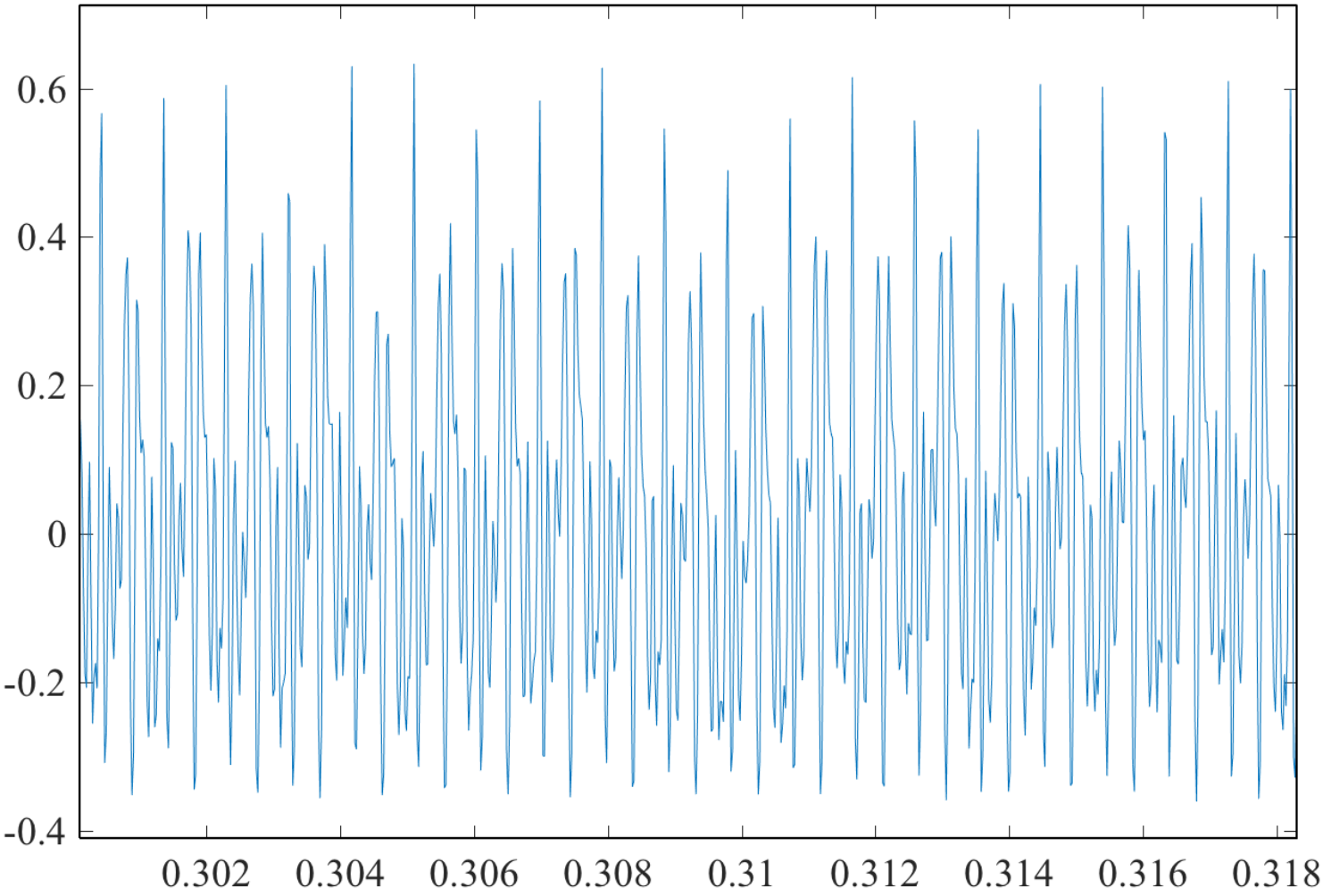
This is Gaussian white noise, filtered with a 50 day lowpass filter.

Apparent structure is due to the interaction of randomness with the filter width. There is nothing phyical about it at all.

Sixth Example



Sixth Example



Observable Features

1. The time series has a very rough appearance.
2. The amplitude of this roughness varies as a function of time.
3. The signal is highly asymmetric, with larger positive amplitudes than negative amplitudes.
4. Amplitude “notches” appear in the positive side, but less so on the negative side.
5. Zooming in, we see the signal roughness is actually composed of *repeated patterns* that are highly non-sinusoidal.

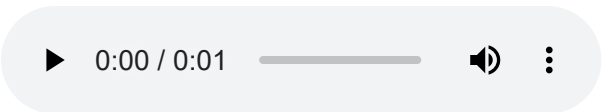
Repeated patterns such as these can be generated by adding up sinusoids having frequencies that are integer multiples of a common frequency, that is, harmonics. This suggests the signal is some kind of vocalization or musical tone.

Observable Features

1. The time series has a very rough appearance.
2. The amplitude of this roughness varies as a function of time.
3. The signal is highly asymmetric, with larger positive amplitudes than negative amplitudes.
4. Amplitude “notches” appear in the positive side, but less so on the negative side.
5. Zooming in, we see the signal roughness is actually composed of *repeated patterns* that are highly non-sinusoidal.

Repeated patterns such as these can be generated by adding up sinusoids having frequencies that are integer multiples of a common frequency, that is, harmonics. This suggests the signal is some kind of vocalization or musical tone.

Sometimes it's helpful to use your ears!

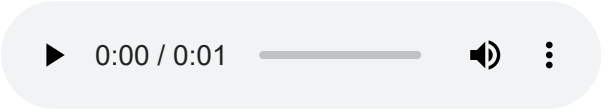


Observable Features

1. The time series has a very rough appearance.
2. The amplitude of this roughness varies as a function of time.
3. The signal is highly asymmetric, with larger positive amplitudes than negative amplitudes.
4. Amplitude “notches” appear in the positive side, but less so on the negative side.
5. Zooming in, we see the signal roughness is actually composed of *repeated patterns* that are highly non-sinusoidal.

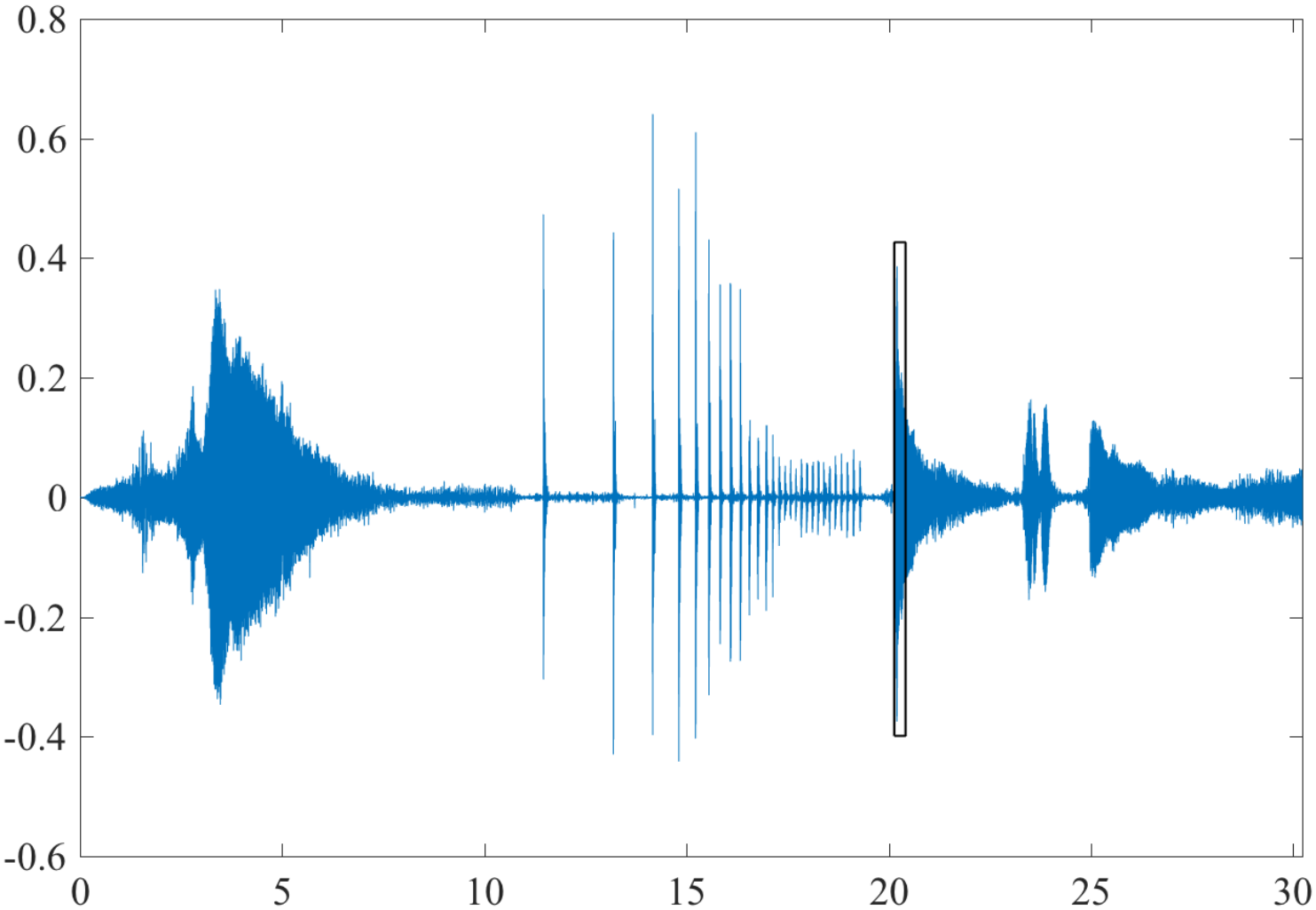
Repeated patterns such as these can be generated by adding up sinusoids having frequencies that are integer multiples of a common frequency, that is, harmonics. This suggests the signal is some kind of vocalization or musical tone.

Sometimes it's helpful to use your ears!

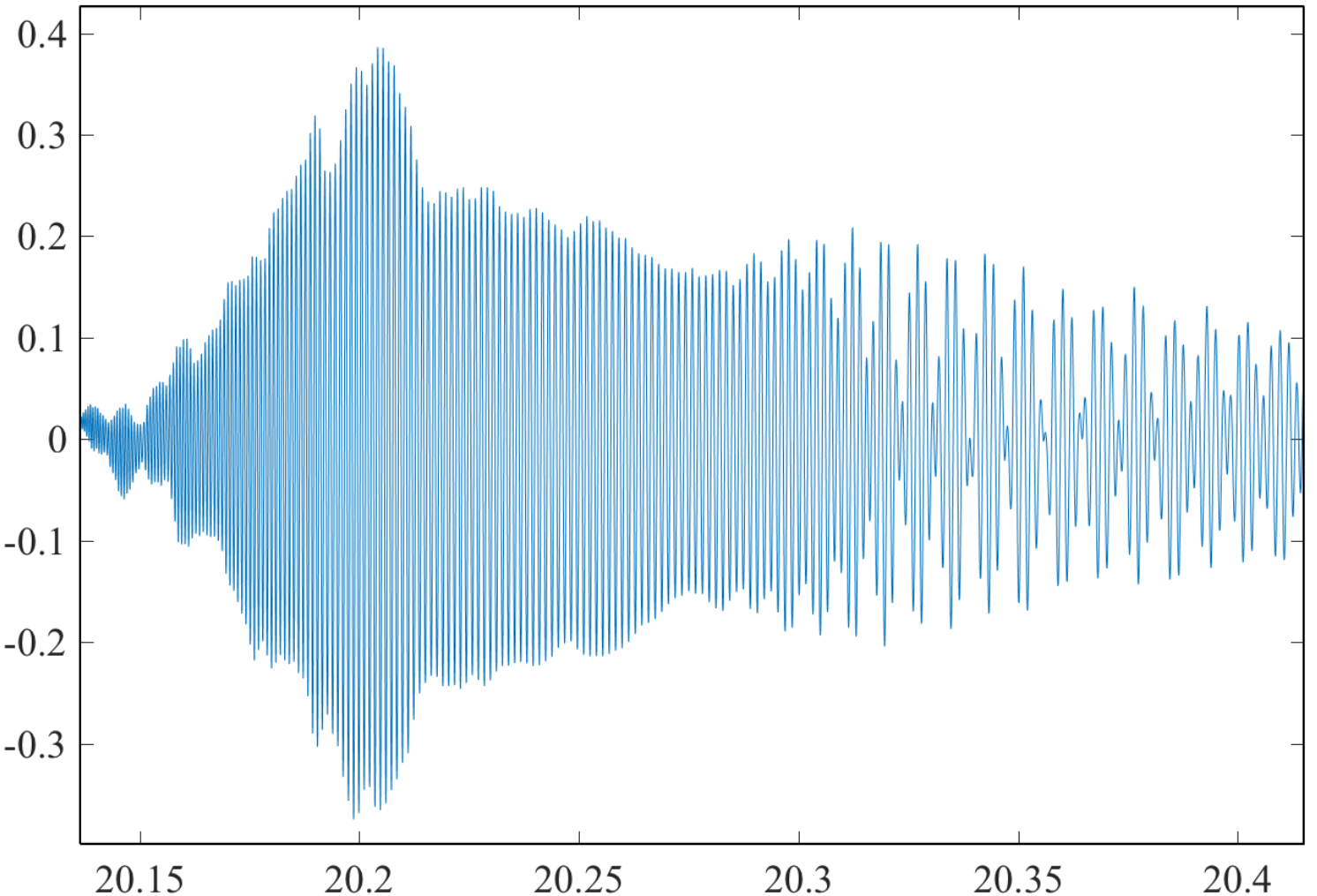


This is an orca call, courtesy of Beam Reach, Seattle.

Seventh Example



Seventh Example



Observable Features

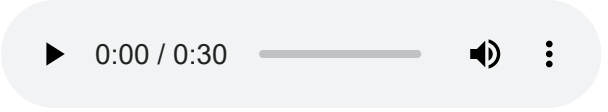
1. The intrinsic noise level appears very low.
2. The time series has two very distinct types of features.
3. The first type has a very dense oscillatory structure, with amplitudes that typically rise rapidly and then fall more slowly.
4. The second type of feature is abrupt spikes. These are very narrow in time, with both an upward portion and a downward portion. These become closer together as time increases.
5. The time series is generally symmetric up/down, but highly asymmetric left/right.
6. Zooming in, we see that the highly oscillatory feature appears to have a frequency that *decreases* with time.

This is a very strange signal. It has features that look like both the seismic signal and the orca vocalization.

Observable Features

1. The intrinsic noise level appears very low.
2. The time series has two very distinct types of features.
3. The first type has a very dense oscillatory structure, with amplitudes that typically rise rapidly and then fall more slowly.
4. The second type of feature is abrupt spikes. These are very narrow in time, with both an upward portion and a downward portion. These become closer together as time increases.
5. The time series is generally symmetric up/down, but highly asymmetric left/right.
6. Zooming in, we see that the highly oscillatory feature appears to have a frequency that *decreases* with time.

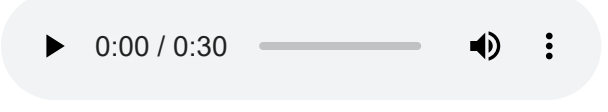
This is a very strange signal. It has features that look like both the seismic signal and the orca vocalization.



Observable Features

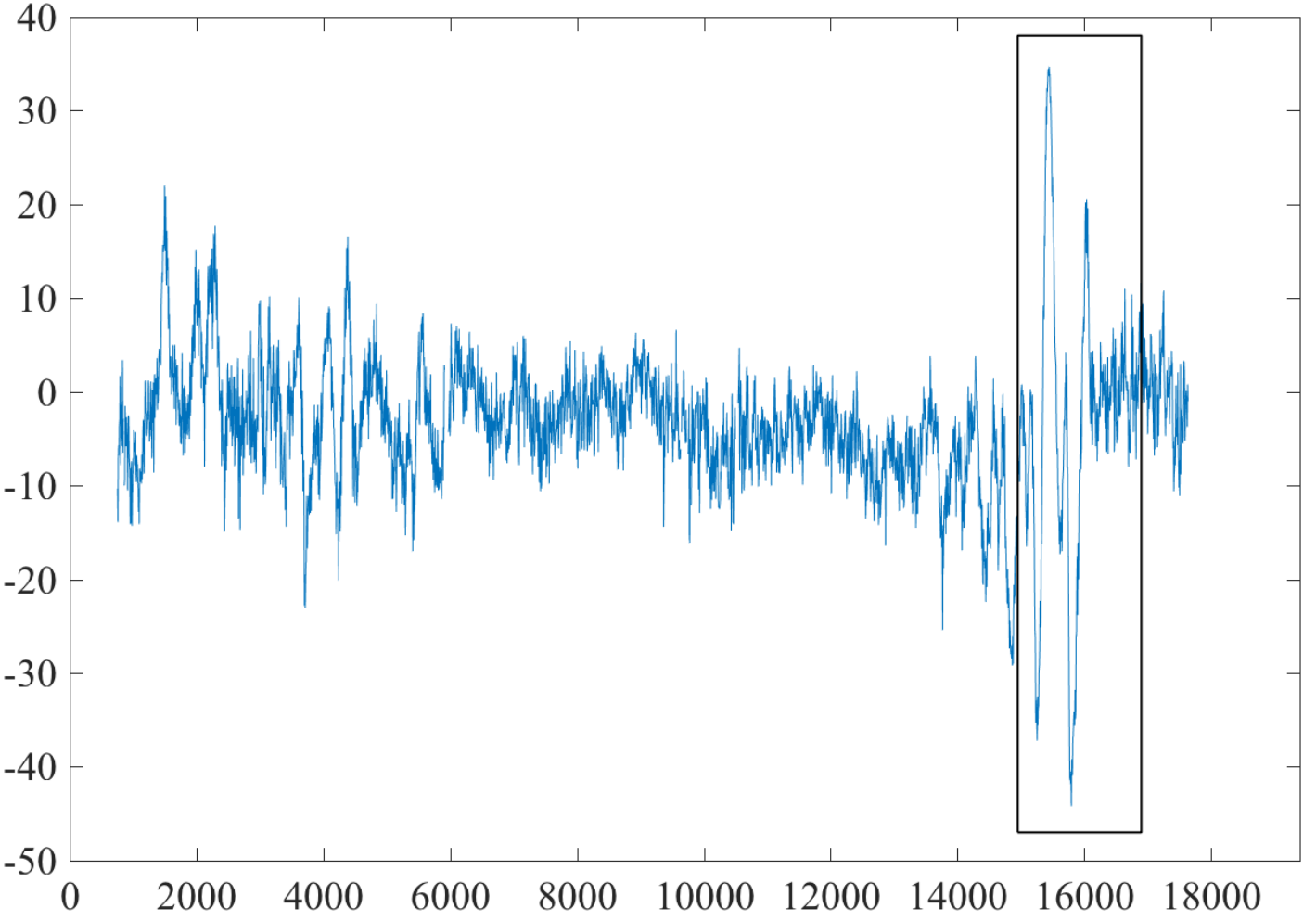
1. The intrinsic noise level appears very low.
2. The time series has two very distinct types of features.
3. The first type has a very dense oscillatory structure, with amplitudes that typically rise rapidly and then fall more slowly.
4. The second type of feature is abrupt spikes. These are very narrow in time, with both an upward portion and a downward portion. These become closer together as time increases.
5. The time series is generally symmetric up/down, but highly asymmetric left/right.
6. Zooming in, we see that the highly oscillatory feature appears to have a frequency that *decreases* with time.

This is a very strange signal. It has features that look like both the seismic signal and the orca vocalization.

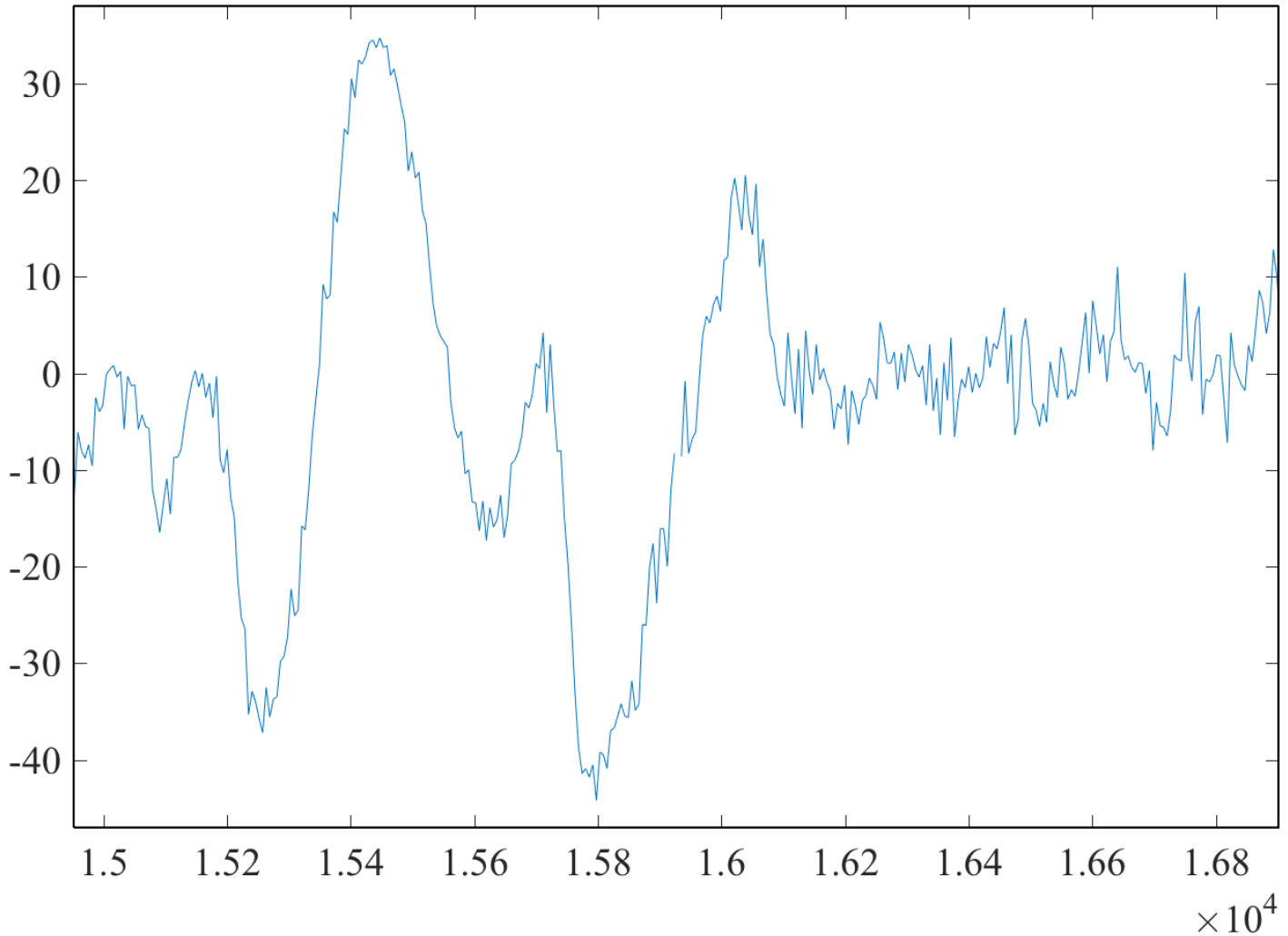


These are Weddell seal calls, courtesy of WeddellSealScience.

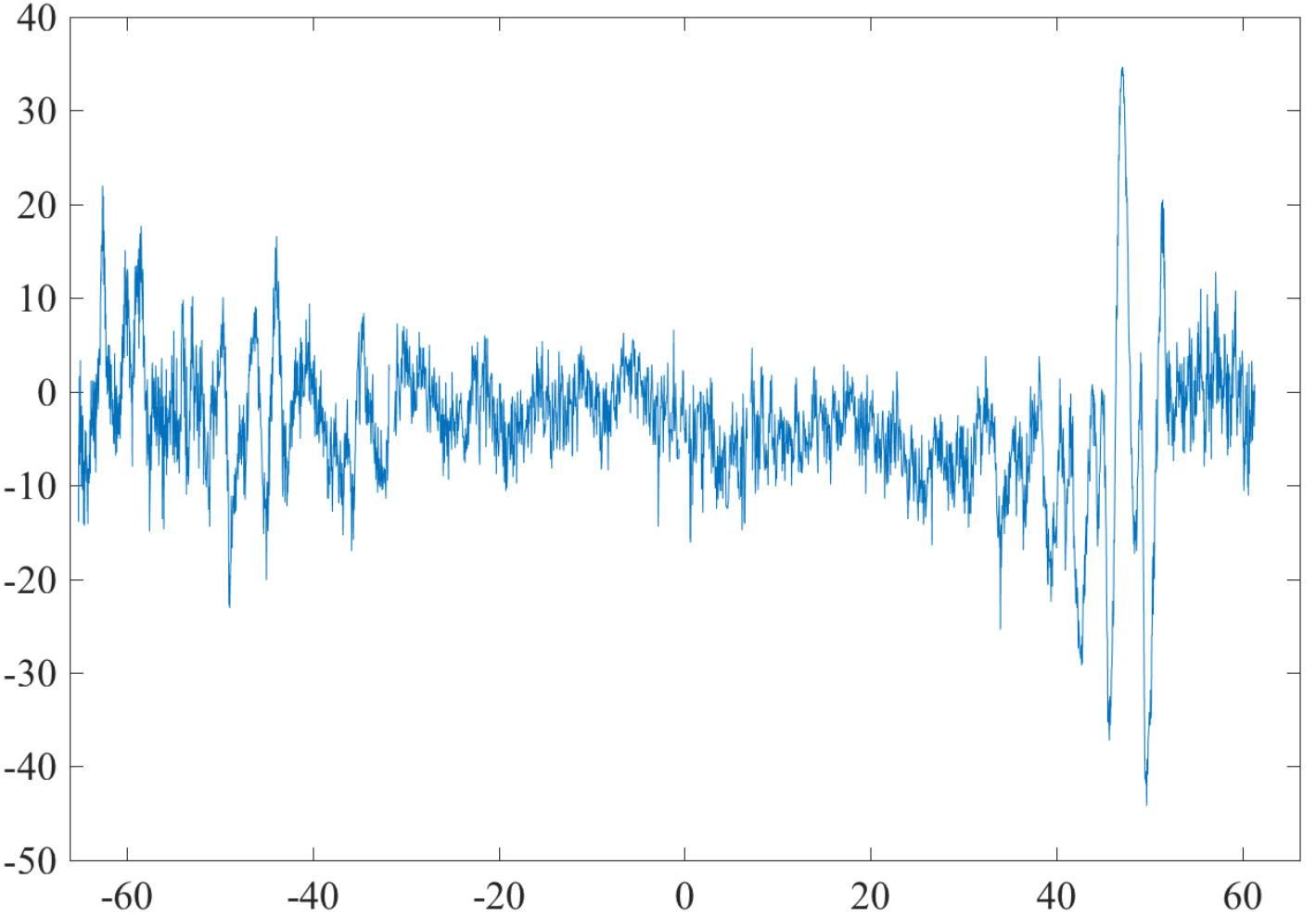
Eighth Example



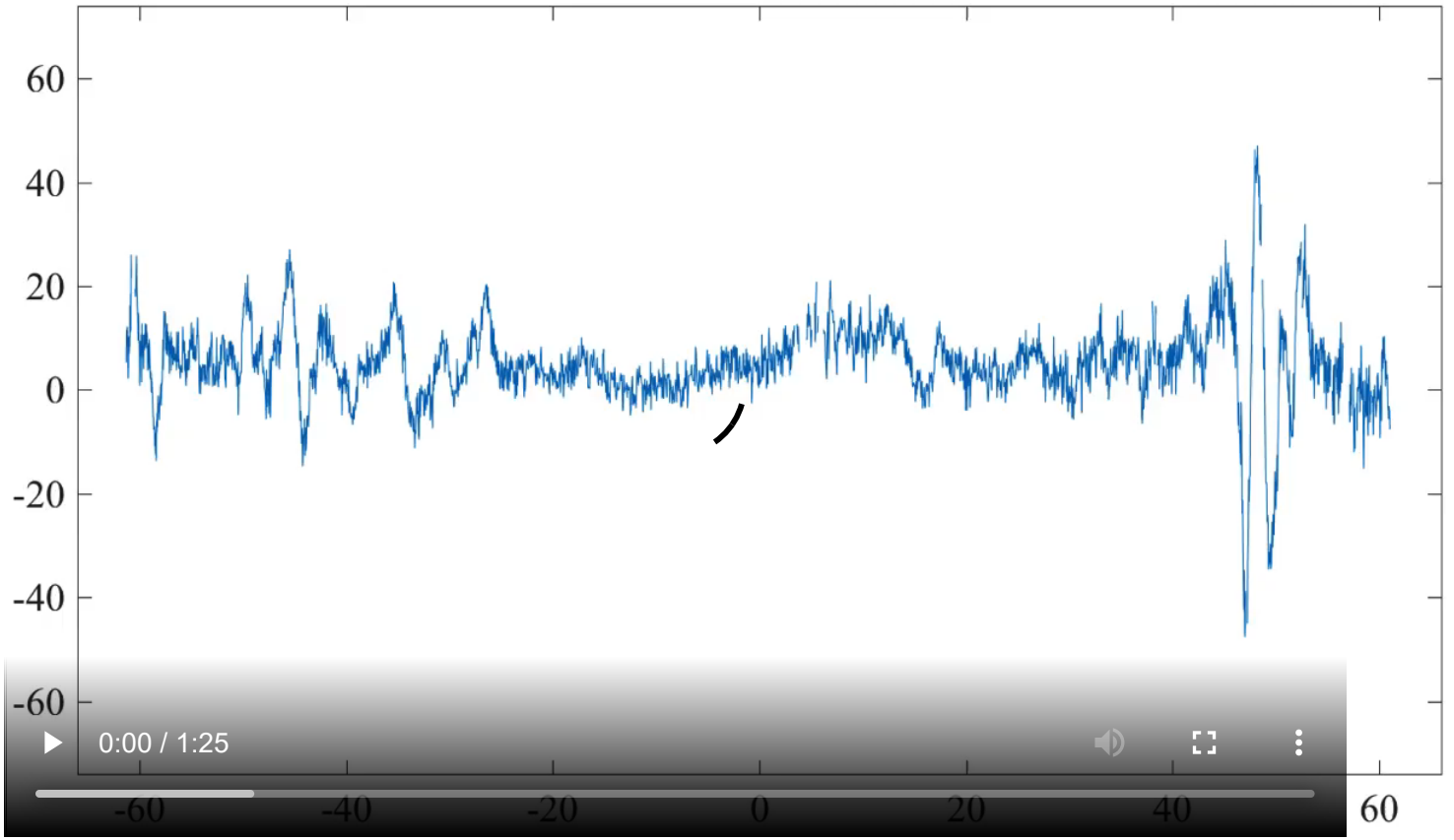
Eighth Example



Eighth Example



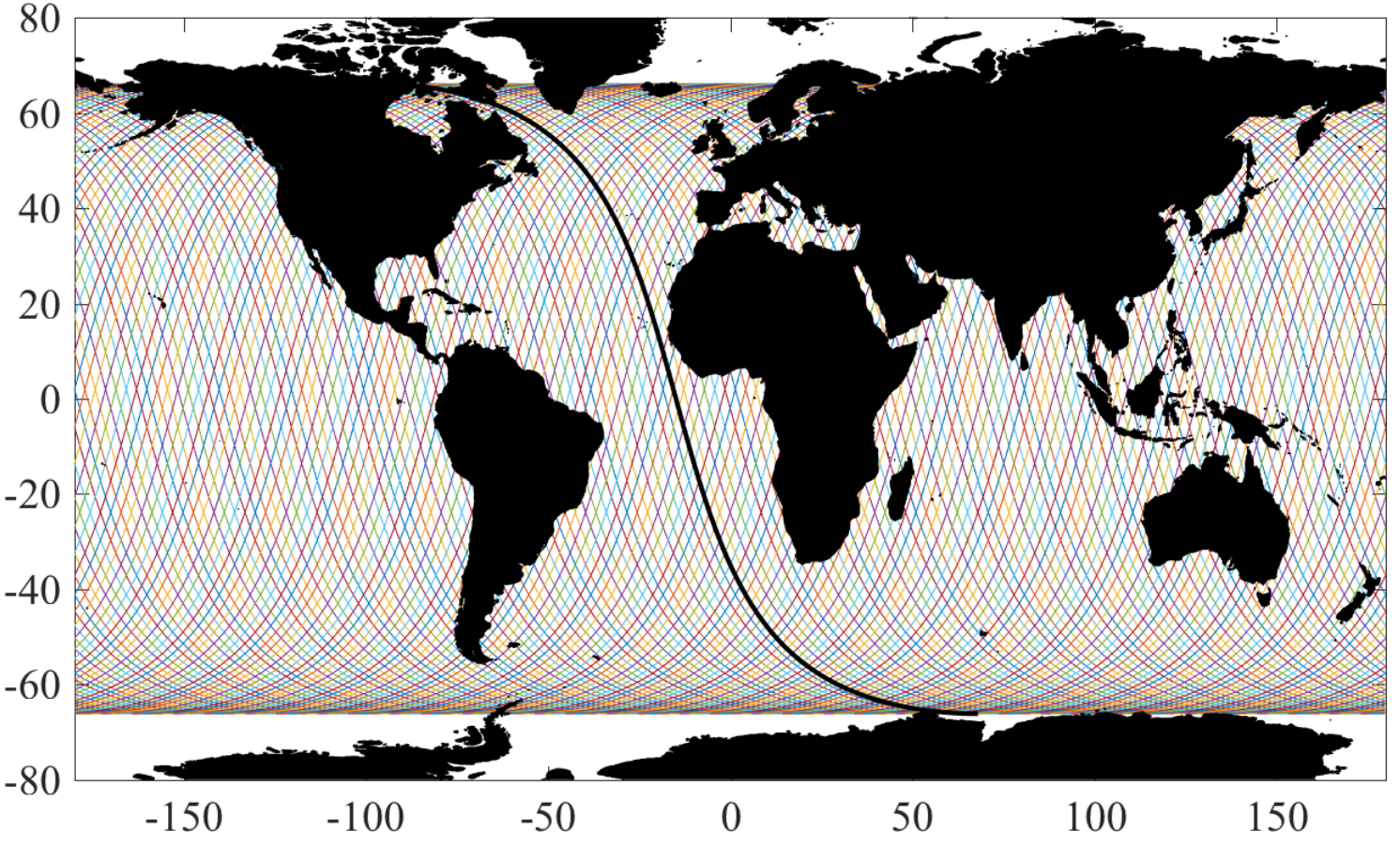
Does it help to see the x-axis?



This is more of the same type of data.

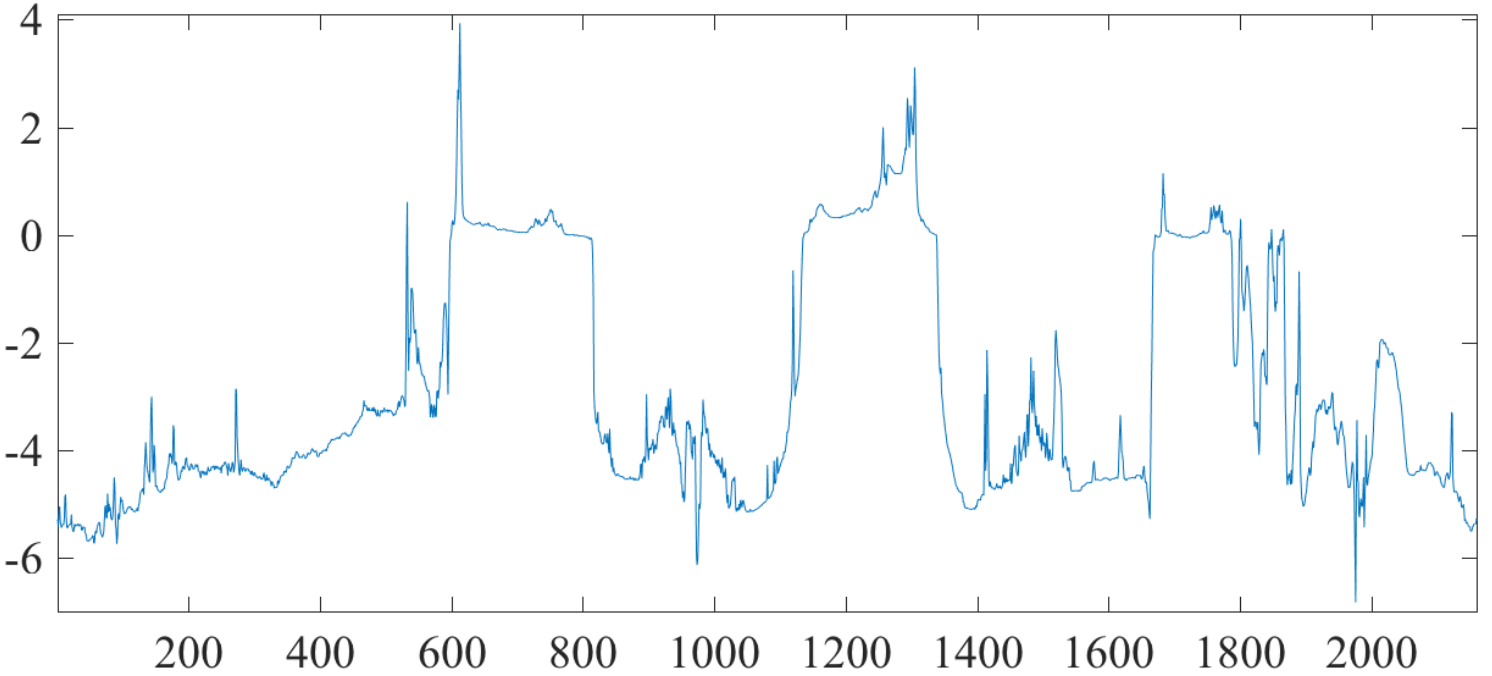
Observable Features

1. The intrinsic noise level appears relatively large compared to the signal, and appears uniformly distributed over all x-locations.
2. The scales of variability vary as a function of the x-axis location. Relatively small x-scales can be seen for $|x| > 55$, intermediate scales in the range $35 < |x| < 55$, and broad scales for $|x| \approx 0$.
3. The largest amplitude variabilty coincides with the band of intermediate scales in the range $35 < |x| < 55$. While there is variability in the vicinity of $x = 0$, the surrounding band $|x| < 20$ is relatively featureless.
4. Large positive excursions appear to be favored over large negative excursions.
5. The pattern is not entirely symmetric in x, as variability in the range $35 \leq |x| < 55$ is typically larger for $x > 0$ than for $x < 0$.
6. In the animation, coherence or persistence of features through several frames is observed.
7. A periodic excursion of missing data is seen on the left-hand side, extending to $x \approx -55$, but not on the right-hand side.

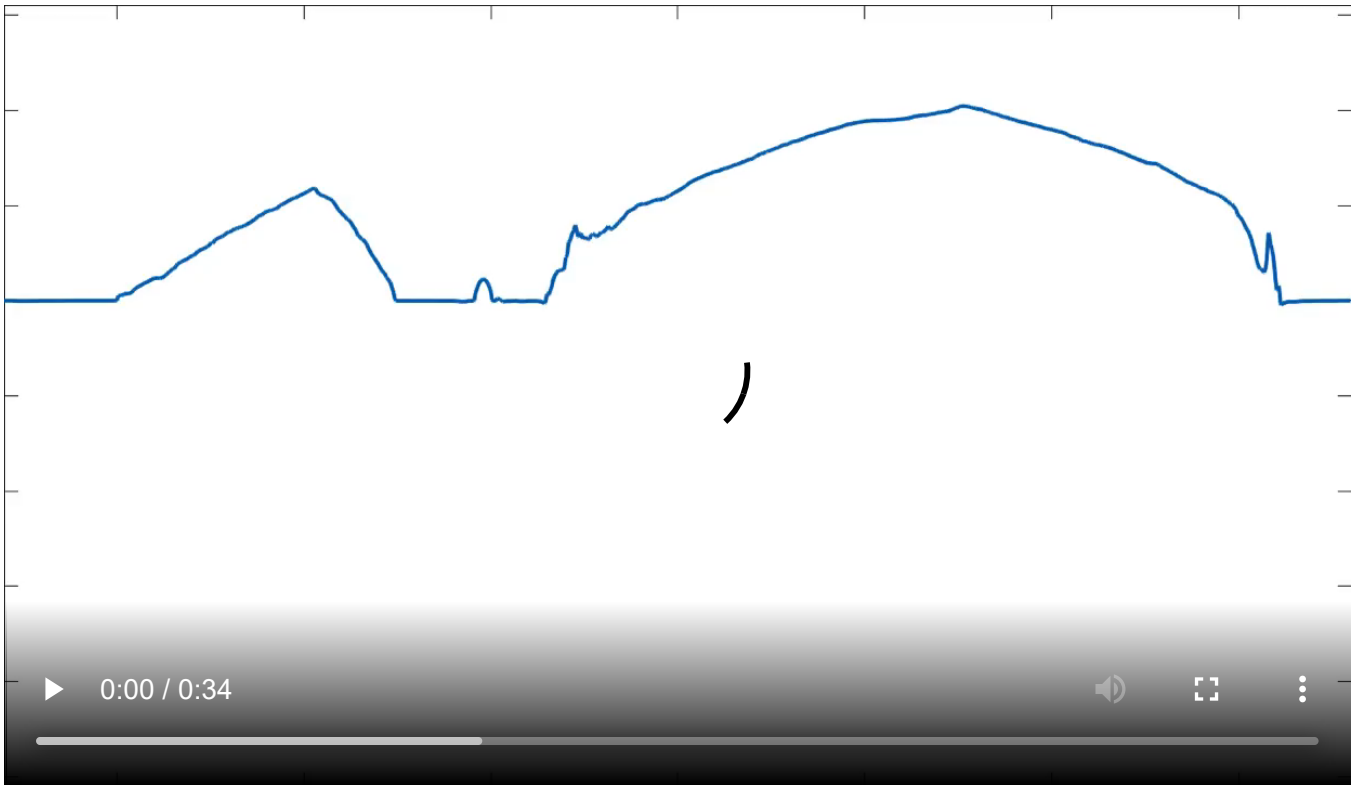


This is Topex/Poseidon/Jason altimetry observed along a single long track, the track highlighted in black, plotted versus latitude. Each animation frame is about 10 days apart.

Ninth Example

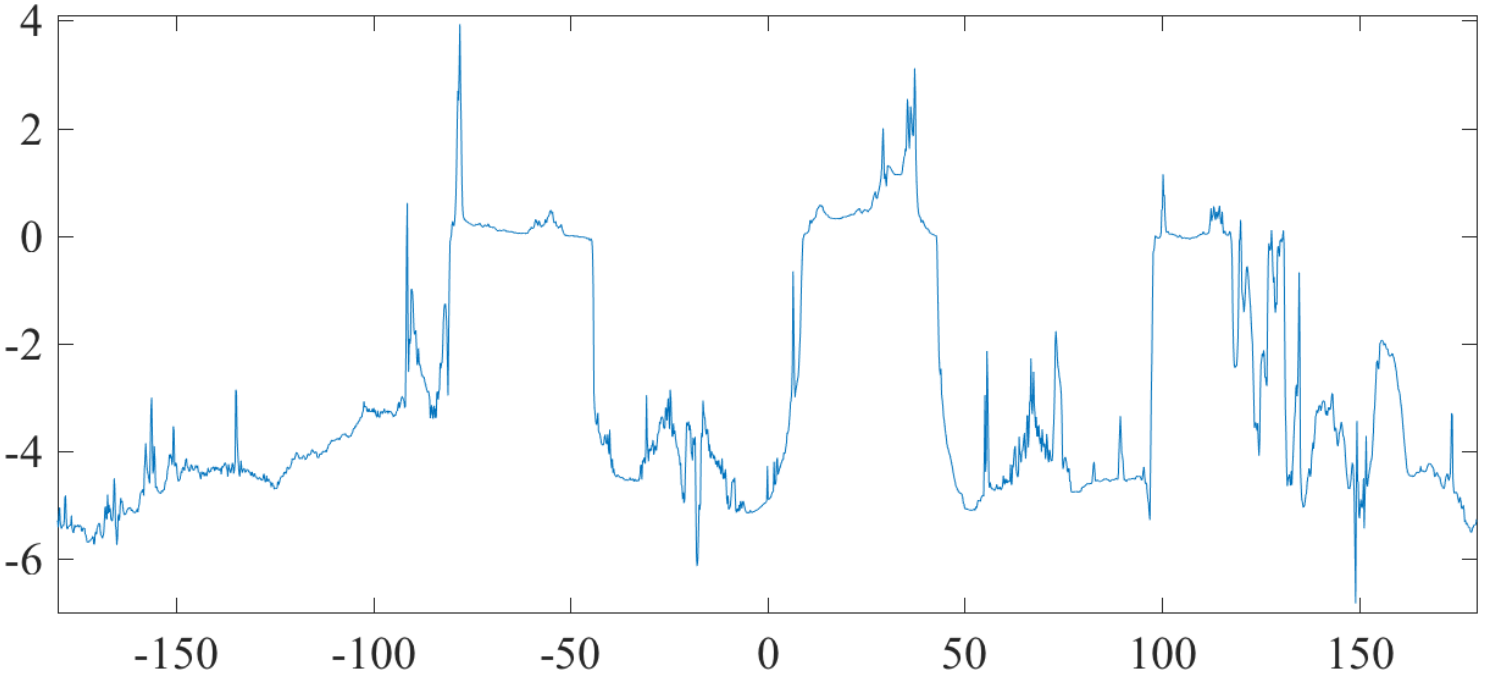


Ninth Example



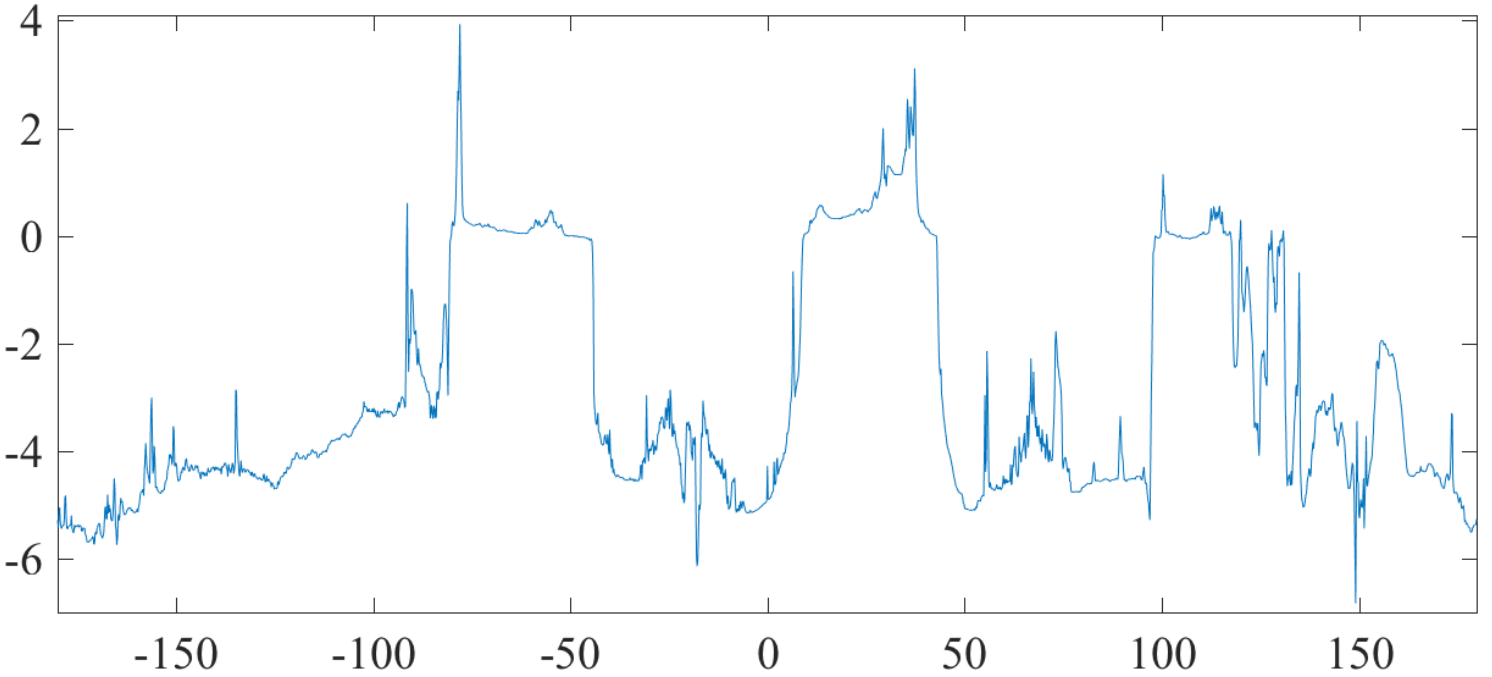
This is more data of the same type. The previous image appears about halfway through.

Ninth Example



Does it help to see the x-axis?

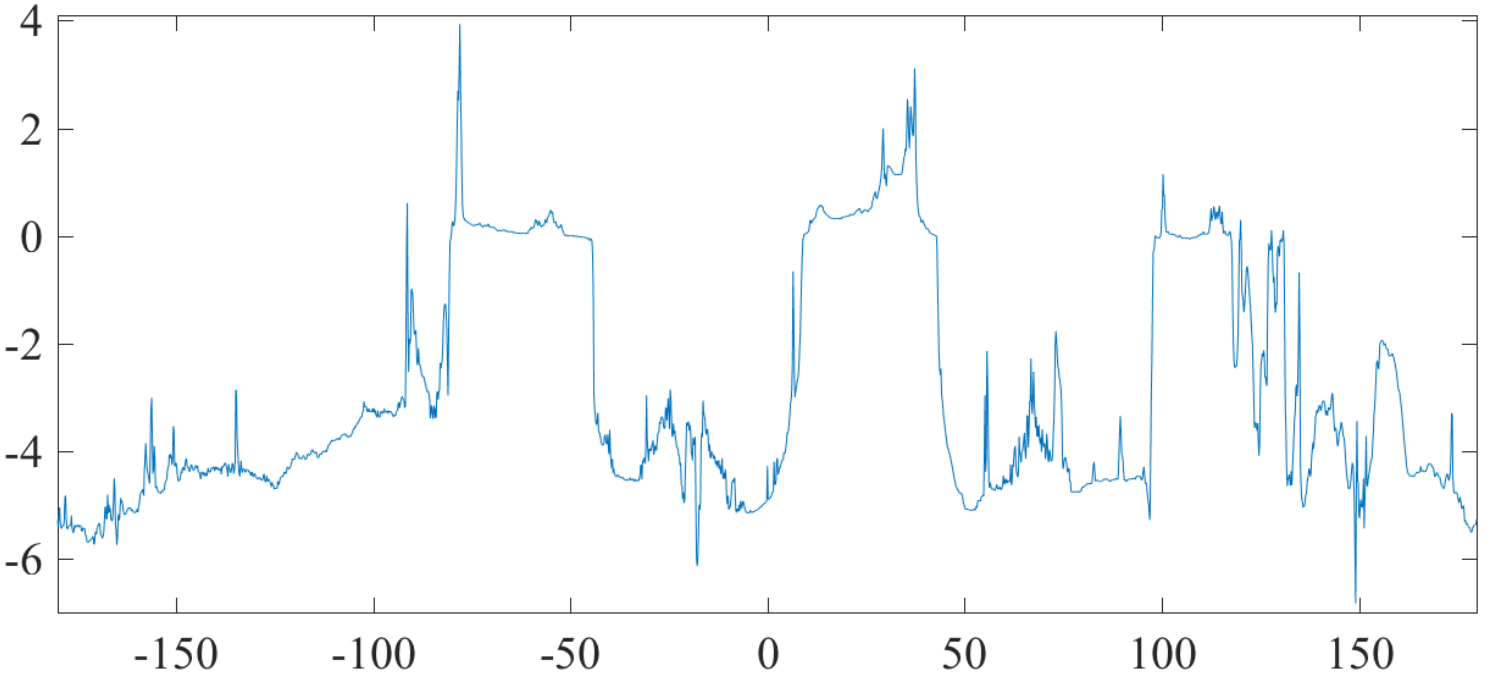
Ninth Example



Does it help to see the x-axis?

The units of the y-axis are kilometers.

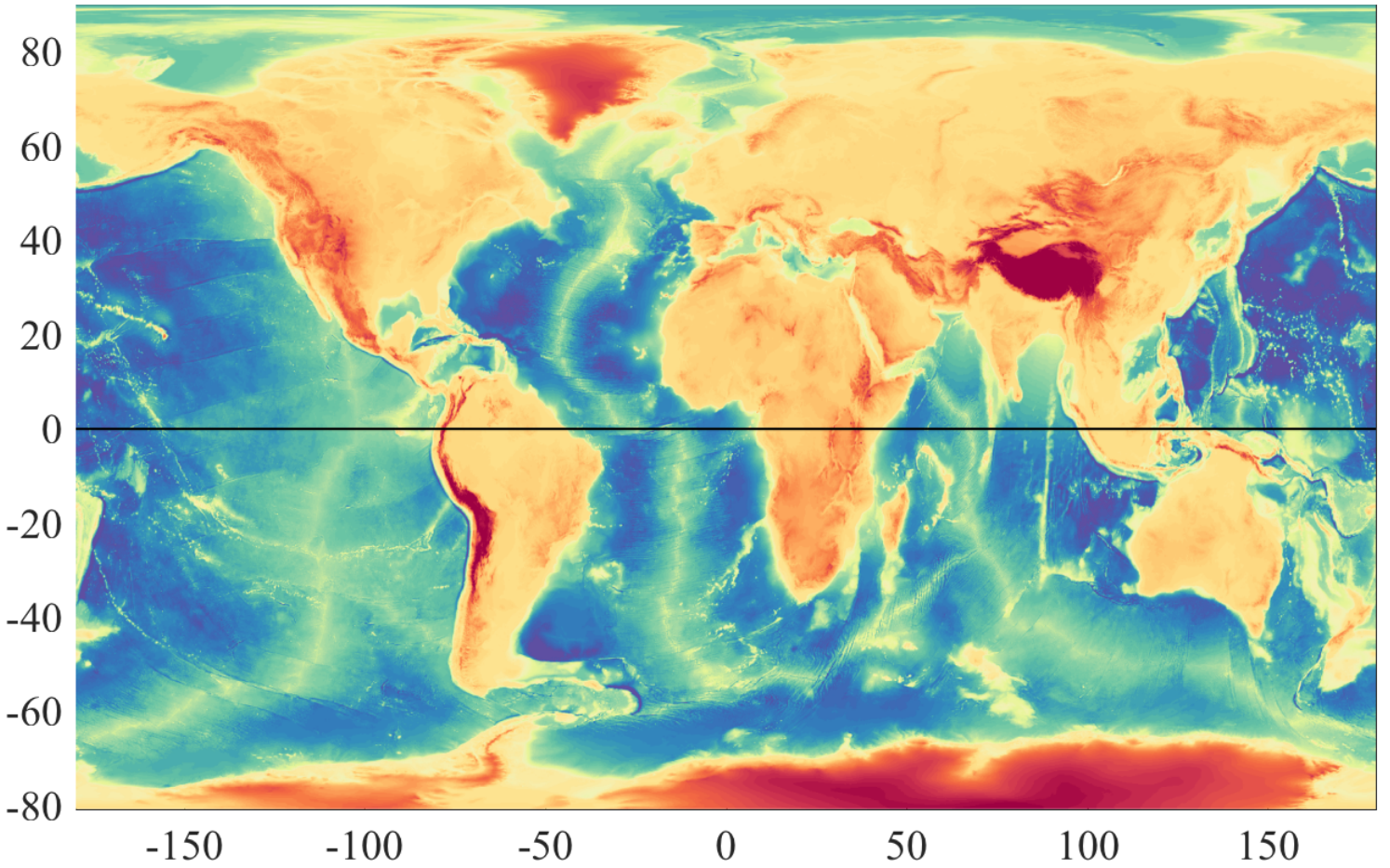
Ninth Example



Does it help to see the x-axis?

The units of the y-axis are kilometers.

The units of the x-axis are degrees.



This is the Earth's topography, sliced along lines of latitude. The animation proceeds from the south to the north.

How to Look at Data

Let the data speak for itself.

How to Look at Data

Let the data speak for itself.

Exercise your powers of observation. How many different features can you see?

How to Look at Data

Let the data speak for itself.

Exercise your powers of observation. How many different features can you see?

Exercise your imagination. What are possible explanations for these different features?

How to Look at Data

Let the data speak for itself.

Exercise your powers of observation. How many different features can you see?

Exercise your imagination. What are possible explanations for these different features?

Consider noise and artifacts. Be aware that features may be spurious, misleading in appearance.

How to Look at Data

Let the data speak for itself.

Exercise your powers of observation. How many different features can you see?

Exercise your imagination. What are possible explanations for these different features?

Consider noise and artifacts. Be aware that features may be spurious, misleading in appearance.

Don't be satisfied with a single interpretation. If you can come up with one interpretation, what is an alternate?

How to Look at Data

Let the data speak for itself.

Exercise your powers of observation. How many different features can you see?

Exercise your imagination. What are possible explanations for these different features?

Consider noise and artifacts. Be aware that features may be spurious, misleading in appearance.

Don't be satisfied with a single interpretation. If you can come up with one interpretation, what is an alternate?

Don't take sides. Don't settle on a single fixed interpretation. Try to hold different points of view at the same time.

How to Look at Data

Let the data speak for itself.

Exercise your powers of observation. How many different features can you see?

Exercise your imagination. What are possible explanations for these different features?

Consider noise and artifacts. Be aware that features may be spurious, misleading in appearance.

Don't be satisfied with a single interpretation. If you can come up with one interpretation, what is an alternate?

Don't take sides. Don't settle on a single fixed interpretation. Try to hold different points of view at the same time.

Don't turn to other tools until you have really looked thoroughly.

Some Questions

What is the overall variability of the time series like? Is it smooth, or rough? Does it change with time? Does variability appear organized at a particular scale or set of scales? Is there “noise”?

Are there excursions? If so, are these symmetric up/down? Are they symmetric front-to-back? Are they uniformly distributed in time?

Are there periodic features? If so, would these be characterized as oscillations? Does the period appear to change in time? Does the oscillation appear regular, like a sinusoid? Are the peaks and valleys symmetric up/down and front-to-back?

Does the sample interval appear sufficient to resolve the variability? Does the duration appear sufficient?

Are there obvious periods of missing data, outliers, or other suspicious features? Where do these tend to occur?

Speed Science!

In this assignment, we first count off into ones and twos. The ones bring up a zoomable image of their dataset, and stay put. The twos circulate throughout the room.

You have five minutes to introduce yourselves, and for the twos to tell the ones what they see in their data.

Note!! It is the person to whom the data does *not* belong who is doing most of the talking! The ones are mostly there to answer questions. Then the bell rings, and all of the twos rotate one position. Sound good? Have fun!

Homework

All homework should be done in a Live Script, as discussed in this afternoon's lab.

Preparing your data:

1. Remove any obvious bad values.
2. If your dataset is not regularly spaced, interpolate it to be so.
3. If your dataset has missing data attend to these through simple linear interpolation; try the `jLab` routine `fillbad`.

Then:

1. Please review the notes.
2. Look at your data using the above idea.
3. Note as many observable features as you can.
4. Comment these in your Live Script.

Also do the homework at the end of the Data Analysis Startup Lab.

If you don't have a mooring dataset you can use {this one}.

The Time Domain

The Sample Interval

We have a sequence of N observations

$$z_n, \qquad n = 0, 1, 2, \ldots N-1$$

which coincide with times

$$t_n, \qquad n = 0, 1, 2, \ldots N-1.$$

The sequence z_n is called a *discrete time series*.

It is assumed that the *sample interval*, Δ, is constant

$$t_n = n\Delta$$

with the time at $n = 0$ defined to be 0. The *duration* is $T = (N-1)\Delta$.

If the sample interval in your data is not uniform, the first processing step is to interpolate it be so.

The Underlying Process

A critical assumption is that there exists some "process" $z(t)$ that our data sequence z_n is a *sample of*:

$$z_n = z(n\Delta), \qquad n = 0, 1, 2, \ldots N-1.$$

Unlike z_n, $z(t)$ is believed to exist for *all times*.

(i) The process $z(t)$ exists in *continuous time*, while z_n only exists at *discrete times*.

(ii) The process $z(t)$ exists for *all past and future* times, while z_n is only available over a certain time interval.

It is the properties of $z(t)$ that we are trying to estimate, *based on* the available sample z_n.

Measurement Noise

In reality, the measurement device and/or data processing probably introduces some artifical variability, termed *noise*.

It is more realistic to consider that the observations contain samples of the process of interest, $z(t)$, *plus* some noise ϵ_n:

$$z_n = z(n\Delta) + \epsilon_n, \qquad n = 0, 1, 2, \ldots N-1.$$

This is an example of the *unobserved components model*. This means that we *believe* that the data is composed of *different components*, but we cannot observe these components individually.

The process $z(t)$ is potentially obscured or degraded by the limitations of data collection in three ways: (i) finite sample interval, (ii) finite duration, (iii) noise.

Because of this, the time series is an *imperfect* representation of the real-world processes we are trying to study.

A Pair of Time Series

In oceanography we often have a *pair* of time series x_n and y_n. Such data is called *bivariate*, meaning that it is consists of two variables.

These may represent horizontal velocity (as in current meters) or displacement (floats or drifters).

Bivariate data can be thought of as a vector having two elements:

$$\mathbf{z}_n = \begin{bmatrix} x_n \\ y_n \end{bmatrix}.$$

The subscript n here refers to n different copies of the vector, *not* to the elements of that vector!

Alternatively, we can also think of this data consisting of a single *complex-valued* time series $z_n \equiv x_n + iy_n$, where $i \equiv \sqrt{-1}$.

Vector and complex notations will both be discussed in detail later.

Time versus Frequency

There are two opposite points of view regarding the time series z_n.

The first regards z_n as being built up as a sequence of discrete values $z_0, z_2, \ldots z_{N-1}$.

This is the domain of *statistics*: the mean, variance, histogram, etc.

When we look at data statistics, generally, the order in which the values are observed *doesn't matter*.

The second point of view regards z_n as being built up of sinusoids: purely periodic functions spanning the whole duration of the data.

This is the domain of *Fourier spectral analysis*.

In between these two extremes is wavelet analysis.

This lecture will focus on what can be done in the time domain.

Time-Domain Statistics

A good place to start is with the very simplest tools. We'll change to x_n and $x(t)$ as this discussion pertains to real-valued data.

The *sample mean* describes a "typical" value:

$$\overline{x} \equiv \frac{1}{N}\sum_{n=0}^{N-1} x_n$$

The *sample variance* gives the spread about the mean:

$$\sigma_x^2 \equiv \frac{1}{N}\sum_{n=0}^{N-1} (x_n - \overline{x})^2$$

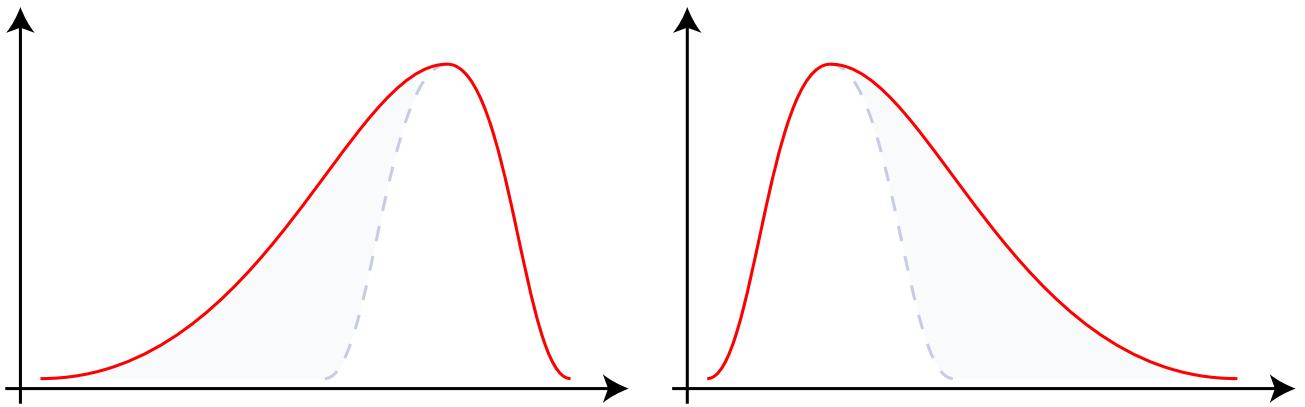
"Sample" here means that it is computed from the observed data, as opposed to being a property of the assumed underlying process $x(t)$.

The mean and variance are called the first two *moments* of the distribution of values associated with the process.

Skewness

The *skewness* describes the tendency for an *asymmetry* between positive excursions and negative excursions:

$$\gamma_x \equiv \frac{1}{\sigma_x^3}\frac{1}{N}\sum_{n=0}^{N-1}(x_n - \overline{x})^3$$



Negative Skew

Positive Skew

Kurtosis

The *kurtosis* is said to either measure *peakedness* (concentration near $\overline{x}$), or a tendency for *long tails* (concentration far from $\overline{x}$):

$$\kappa_x \equiv \frac{1}{\sigma_x^4}\frac{1}{N}\sum_{n=0}^{N-1}(x_n - \overline{x})^4$$

Actually, it measures both. Kurtosis is a measure of the spread of x_n about the *two points* $\overline{x} \pm \sigma_x$. This can happen *either* for peakness *or* for long tails! *See Moors (1986), "The Meaning of Kurtosis".*

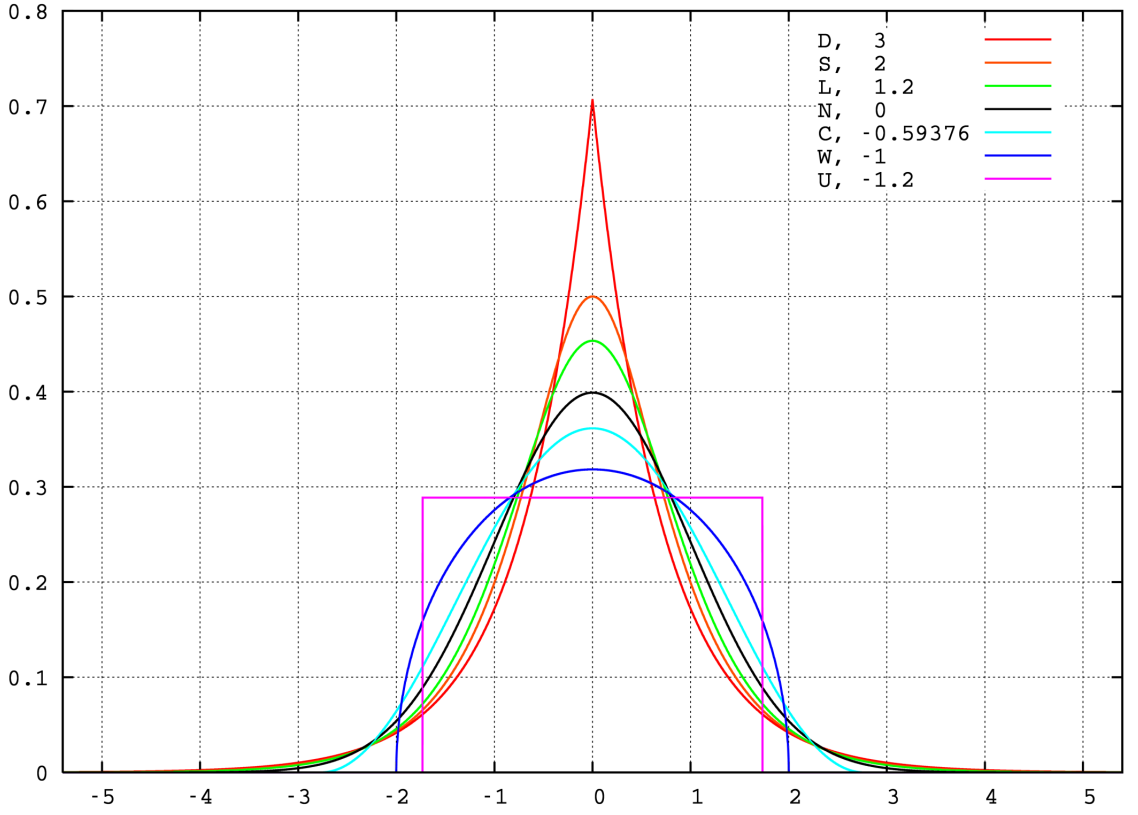
Because the value of kurtosis for a Gaussian process can be shown to be equal to 3, one sometimes encounters the *excess kurtosis*

$$\tilde{\kappa}_x \equiv \kappa_x - 3.$$

Values of $\tilde{\kappa}_x > 0$ mean the data is *more kurtotic*—peaked or long-tailed—than a Gaussian, while $\tilde{\kappa}_x < 0$ means it is less so.

Illustration of Kurtosis

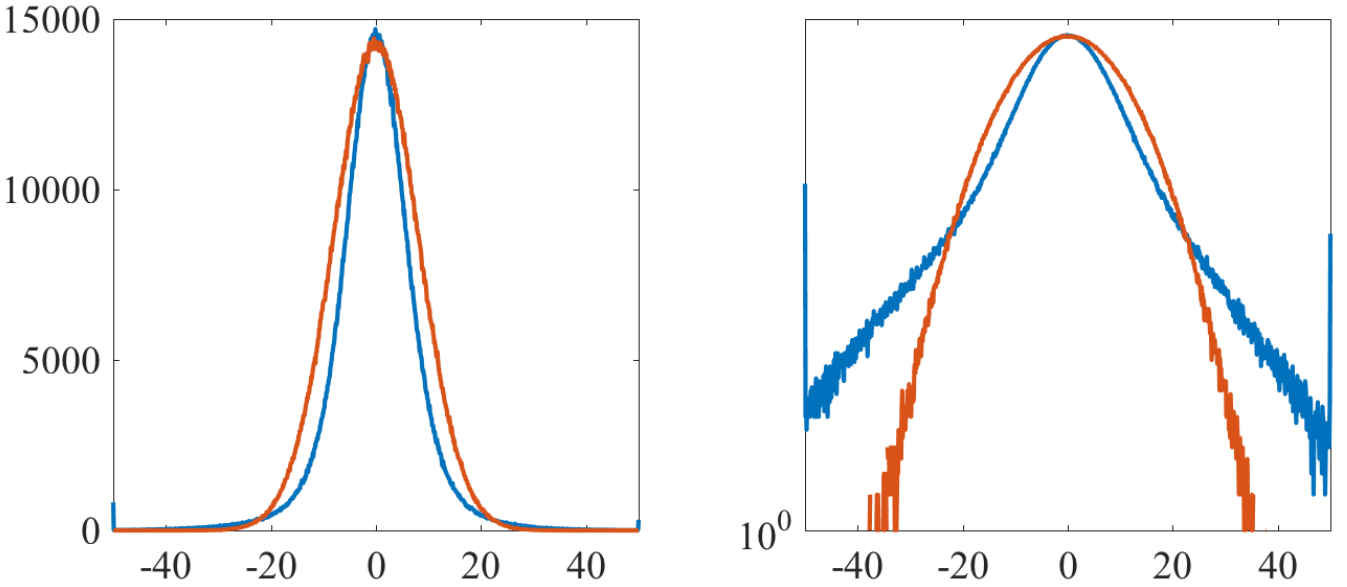
Distributions corresponding to different values of excess kurtosis.



Positive excess kurtosis corresponds to long tails and peakedness.

Histogram

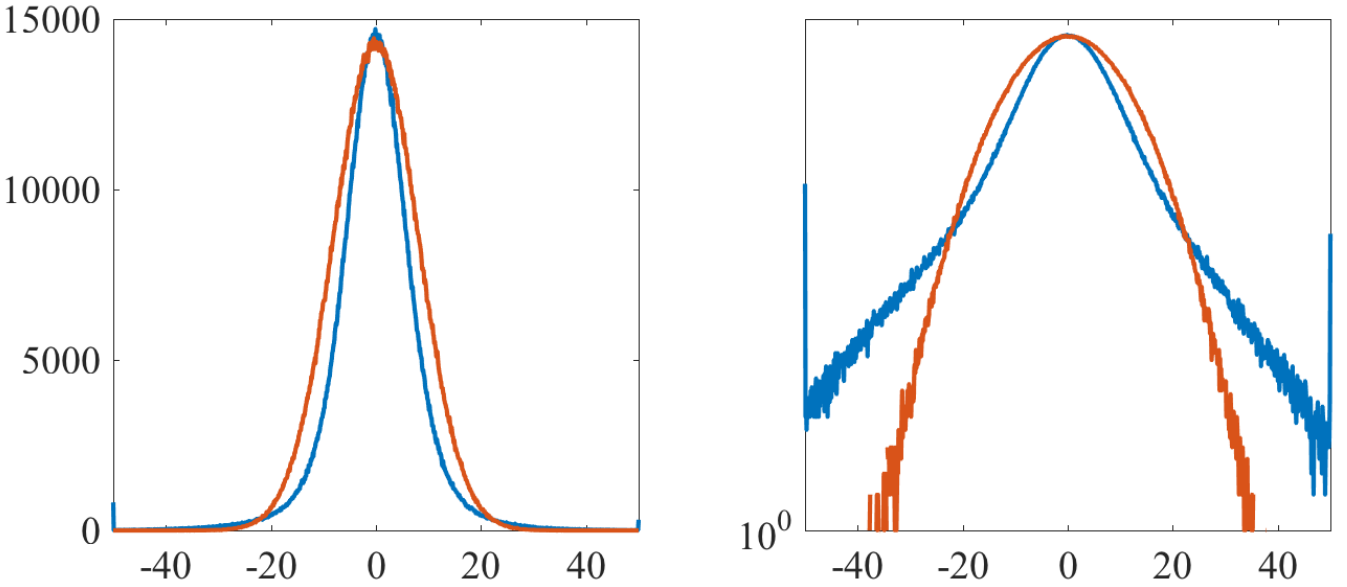
The mean, variance, skewness, and kurtosis describe aspects of the *histogram*: the observed distribution of data values.



Here is the histogram of *all* SSH values from long altimeter track (blue), versus Gaussian noise having the same variance (orange).

Histogram

The mean, variance, skewness, and kurtosis describe aspects of the *histogram*: the observed distribution of data values.



Here is the histogram of *all* SSH values from long altimeter track (blue), versus Gaussian noise having the same variance (orange).

Simple Smoothing

One of the most effective ways to process a time series is with a simple smoothing.

Let g_m be a length M sequence, where M is *odd*, defined for

$$-(M-1)/2, \ldots, -2, -1, 0, 1, 2, \ldots, (M-1)/2.$$

Note that we define g_m to be centered on $m = 0$, instead of running between 0 and $M-1$.

A *smoothed* version of the discrete time series z_n is defined as

$$\tilde{z}_n = \sum_{m=-(M-1)/2}^{(M-1)/2} z_{n-m}\, g_m$$

where g_m is called the *filter* or the *smoothing window*. It is also useful to examine the *residuals* from the original, $\breve{z}_n \equiv z_n - \tilde{z}_n$.

Simple Smoothing Example

An example of simple smoothing is a *running mean*. A five-point running mean is given by:

$$\tilde{z}_n = \frac{1}{5}\left[z_{n-2} + z_{n-1} + z_n + z_{n+1} + z_{n+2}\right].$$

This is expressed by the filtration equation

$$\tilde{z}_n = \sum_{m=-(M-1)/2}^{(M-1)/2} z_{n-m}\, g_m$$

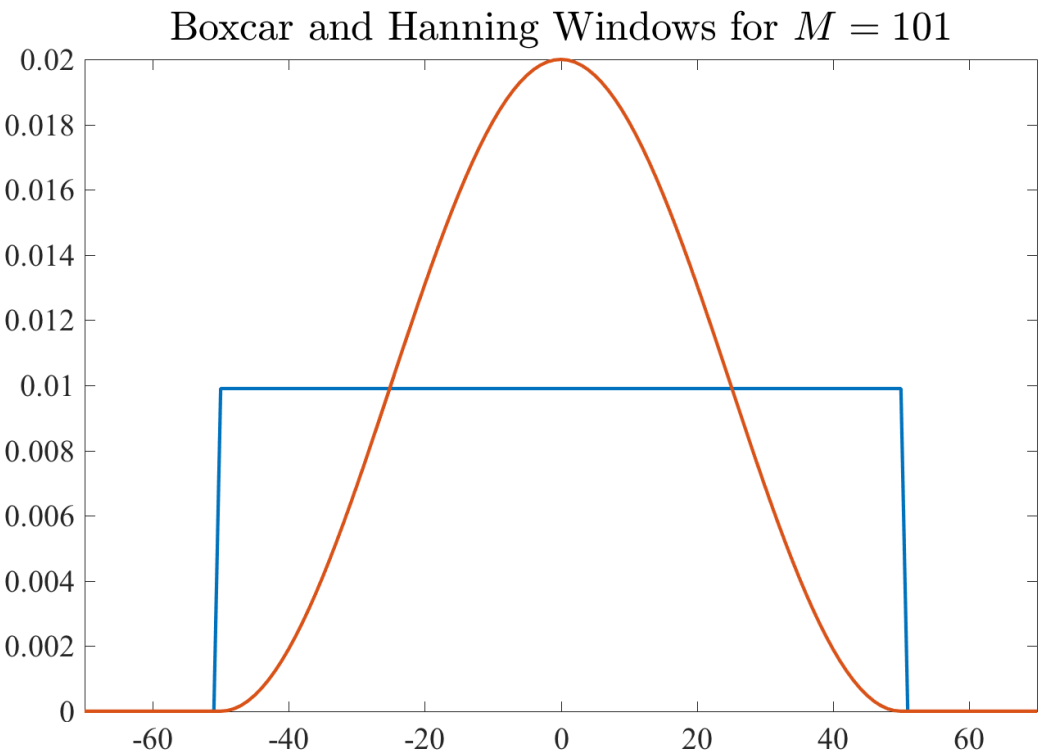
with the choice

$$g_m = 1/5, \qquad m = -2, -1, 0, 1, 2.$$

The simplest choice of filter is $g_m = 1/M$, a constant over the M points. Then the filtration defines an M-point running mean.

Choice of Filter

The running mean filter $g_m = 1/M$ is called the *boxcar* or *rectangle function*. Another popular choice is the *Hanning window*.



The Hanning window is just a half-period of a cosine, offset.

How to Choose a Filter

The goal of simple smoothing is to separate relatively “fast” from relatively “slow” variability.

Many functions can be used as smoothing filters. However, for a first look at the data, the details of the filter are not so important.

The important thing is to define a sensible *weighted average*.

The boxcar filter has sharp “edges” that can lead to artifacts, as we will see later. Also, the boxcar is highly distributed, and doesn't place emphasis on the “present time” compared to nearby times.

For these reasons, the Hanning window is sometimes more appropriate for simple smoothing.

In `jLab` simple smoothing is carried out with `vfilt`.

What to do at Endpoints?

Smoothing runs into a difficulty near the endpoints of z_n:

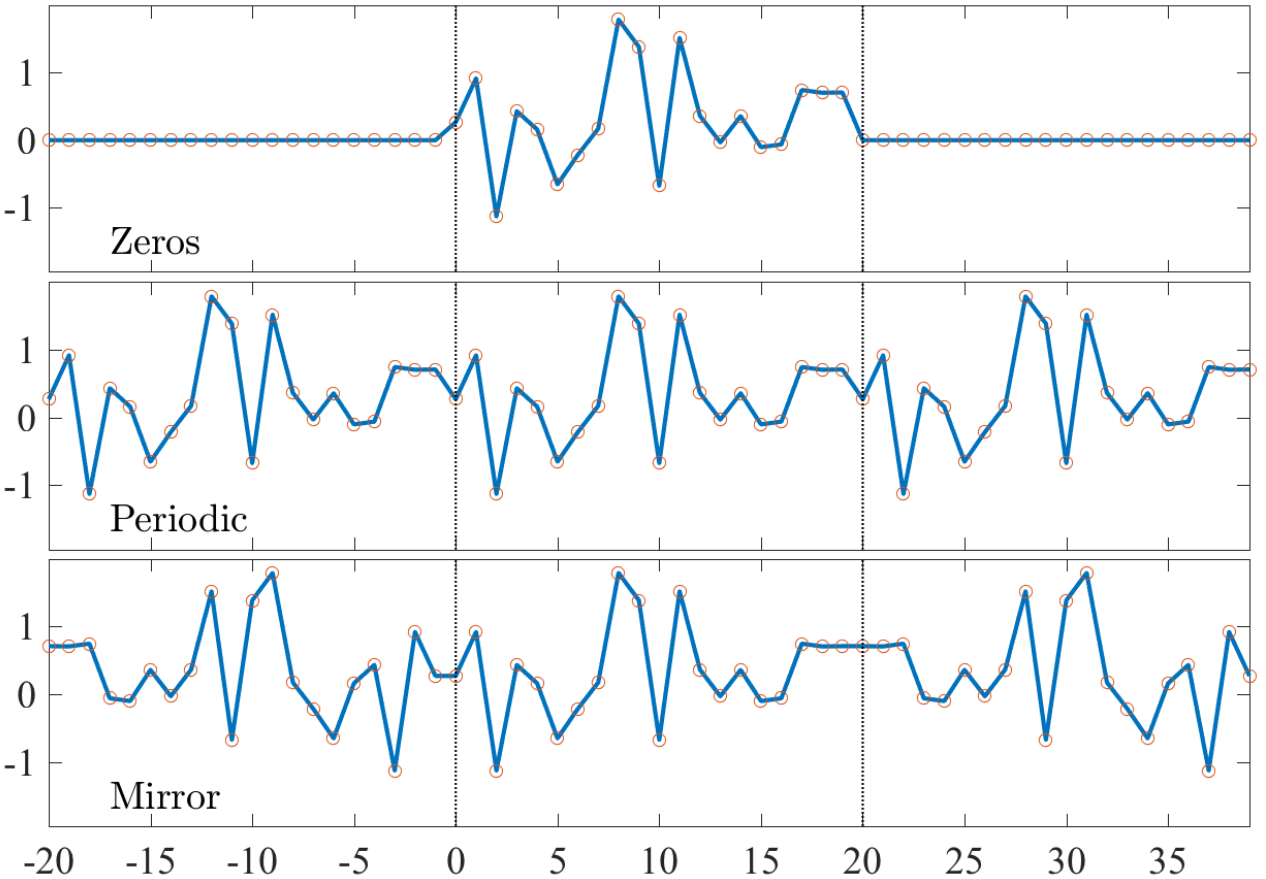
$$\tilde{z}_n = \sum_{m=-(M-1)/2}^{(M-1)/2} z_{n-m}\, g_m.$$

When we are within a filter half-width $(M-1)/2$ of the beginning or end of z_n, the filter "falls off" the end of the data.

Some choice must be made in order to have the smoothed version $\tilde{z}_n$ of the data be well defined. There are five common choices.

1. **Truncate**: Omit affected points, such that the length of $\tilde{z}_n$ will be about M points *less than* the length of z_n.
2. **NaNs**: Replace these with NaNs or *indeterminate* values.
3. **Zeros**: Set z_n equal to zero for $n \leq 0$ or $n \geq N-1$.
4. **Periodic**: Make z_n periodic by wrapping around the ends.
5. **Mirror**: Reflect z_n about its beginning and also about its end.

Endpoint Illustration



The *mirror* condition generally leads to the fewest “edge effects”, especially when the data is nonstationary or has a linear trend.

Summary

This lecture has focused on

Summary

This lecture has focused on

• Introducing the concepts of *discrete sampling*, *sample interval*, *measurement noise*, and the *underlying process*.

Summary

This lecture has focused on

• Introducing the concepts of *discrete sampling*, *sample interval*, *measurement noise*, and the *underlying process*.

• Defining the first four *moments*—mean, variance, skewness, and kurtosis—as well as the *histogram*.

Summary

This lecture has focused on

• Introducing the concepts of *discrete sampling*, *sample interval*, *measurement noise*, and the *underlying process*.

• Defining the first four *moments*—mean, variance, skewness, and kurtosis—as well as the *histogram*.

• Discussing *simple smoothing* and details of its implementation.

Summary

This lecture has focused on

• Introducing the concepts of *discrete sampling*, *sample interval*, *measurement noise*, and the *underlying process*.

• Defining the first four *moments*—mean, variance, skewness, and kurtosis—as well as the *histogram*.

• Discussing *simple smoothing* and details of its implementation.

My experience is that *looking at data* together with *statistics* and *simple smoothing* is maybe 50% of analyzing time series!

There are more sophisticated tools that can often, but not always, be very useful in unlocking the potential of the data.

However, learning how to make use of these takes a lot more work!

To be continued...

Three Cases

In general we have three types of data.

1. Gridded data
2. Dense irregular data
3. Sparse irregular data

We have to approach these differently.

Gridded data

With gridded data we can fruitfully analyze using two-dimensional statistics by directly averaging in different directions along a 2D, 3D, or N-D "cube" of data.

Averaging can be done quickly, without explicit loops, for such data. Then we imagine turning the cube in different directions and averaging along different axes.

It is often useful to split time into two dimensions. For example,

$$\text{lat} \times \text{lon} \times \text{time}$$

can be reorganzied to become

$$\text{lat} \times \text{lon} \times \text{time of year} \times \text{different years}.$$

So then averaging over the 4th dimension creates a composite year, while taking the standard deviation gives the year-to-year variability.

✧ For this, `reshape` and `permute` are useful.

Gridded data

If the data is so big that we can't load it into memory all at one, then we can average by aggregating: loading one time slice in at a time, adding to a running total, then dividing by the number of slices at the end.

The variance can also be computed in this way using formulas like

$$\overline{(x-\overline{x})^2} = \overline{x^2} - \overline{x}^2.$$

So we can create the variance by creating aggregated averages of both $\overline{x}$ and $\overline{x^2}$. The same applies to higher-order moments.

We can also use `twodhist` and `twodstats` to examine distributions and averages in parameter space, e.g. the distribution of sea surface temperature vs. sea level pressure for all locations and all times.

If we find find a pattern in parameter space, it can be quantified using regression analysis, if desired.

Dense irregular data

For irregularly sampled (non-gridded) data, we cannot directly average, but we can use `twodhist` and `twodstats` to examine distributions and averages is sensible.

If there are enough data points such these histograms are sufficiently “filled in”, we will call data “dense”.

With dense irregular data, as with gridded data, we imagine the dataset to be a large multivariate “cloud”, e.g.

$$\text{lat} \times \text{lon} \times \text{time} \times \text{temperature} \times \text{pressure}$$

and then we use the distributional analysis to slice it in different ways, looking for patterns.

It is often very useful to have two-dimensional statistics with quantities of different units on the x- and y-axes, e.g. latitude $\times$ time. Then we can look at distributions, means, and standard deviations, etc. in this plane.

Sparse data

Sparse simply means there is not very much data. Practically speaking, our approach must be different when there is not enough data to fill in a histogram.

In this case, we learn a lot by employing scatter plots making creative use of size and/or color of symbols.

In Matlab this is doen using `scatter`.

Statistics Example

As an example of how to use time-domain statistics, we will look at a numerical model of the Gulf of Mexico.

First consider the mean and standard deviation of the velocity $\mathbf{z}_n(x,y) \equiv [u_n(x,y) \;\; v_n(x,y)]^T$,

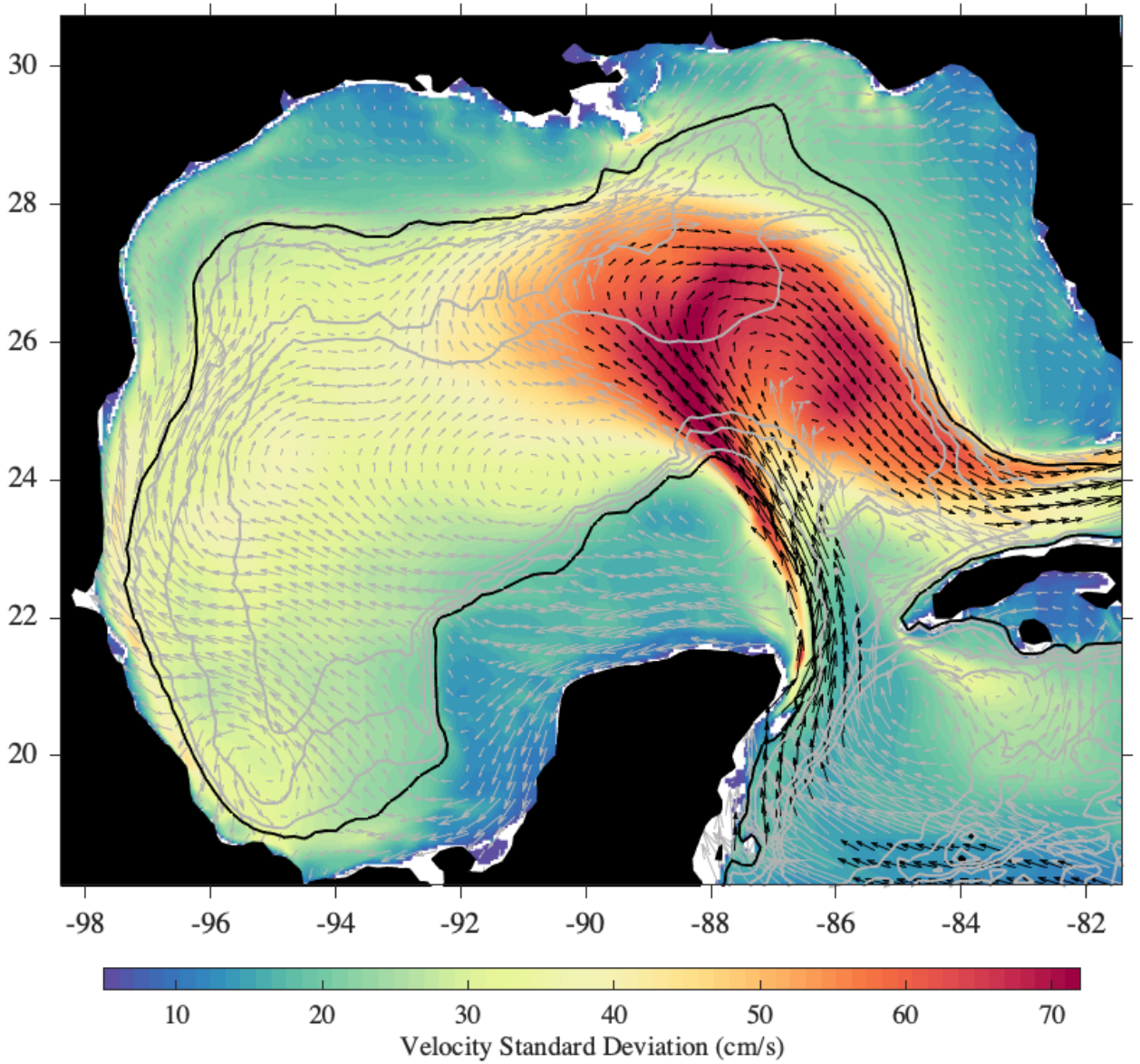
$$\overline{\mathbf{z}}(x,y) \equiv \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{z}_n, \qquad \sigma^2(x,y) \equiv \frac{1}{N}\sum_{n=0}^{N-1}\left(\mathbf{z}_n - \overline{\mathbf{z}}\right)^T\left(\mathbf{z}_n - \overline{\mathbf{z}}\right)$$

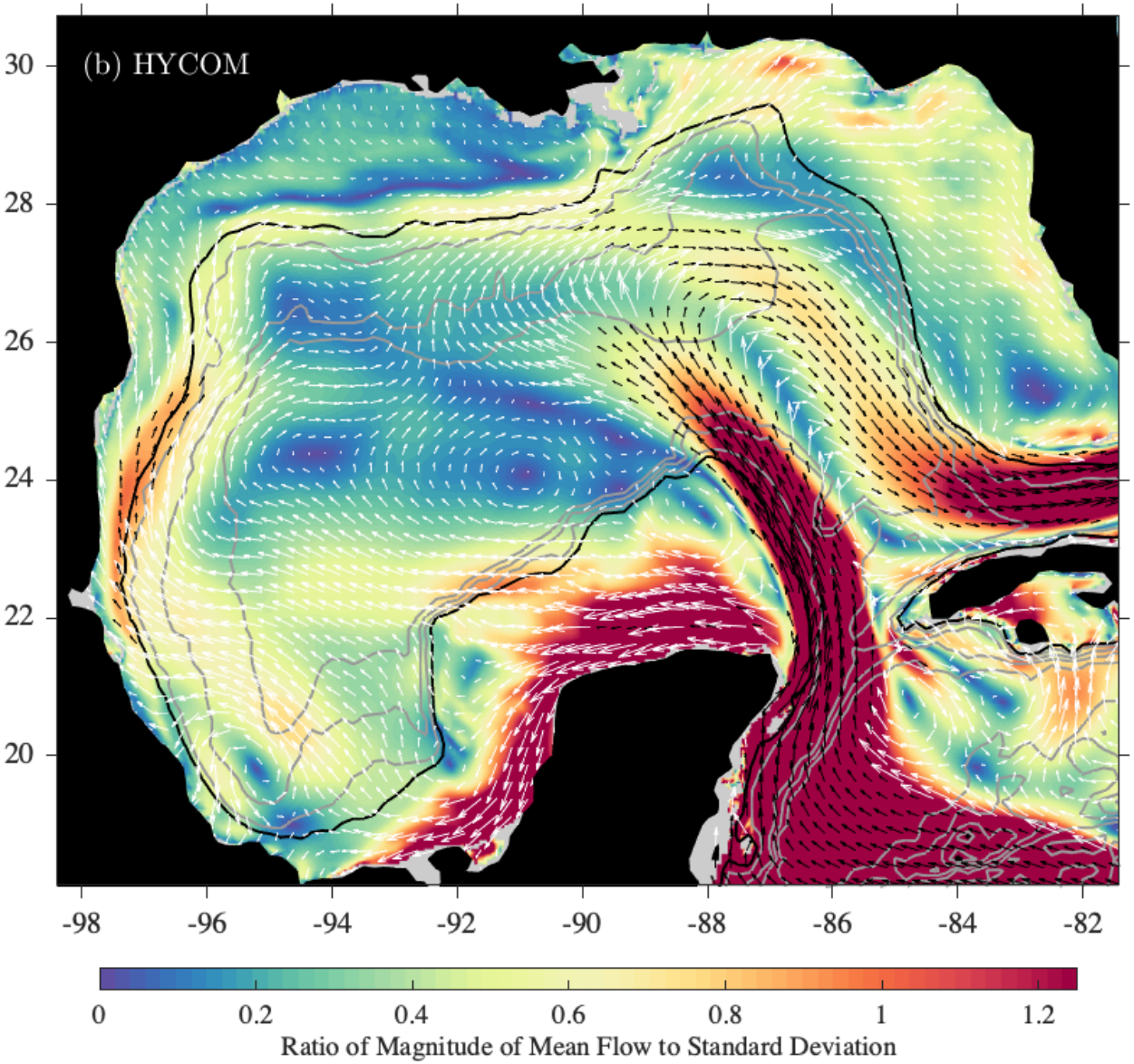
and the ratio of the mean flow magnitude to the standard deviation

$$\frac{\|\overline{\mathbf{z}}(x,y)\|}{\sigma(x,y)} = \frac{\sqrt{\overline{u}^2(x,y) + \overline{v}^2(x,y)}}{\sigma(x,y)}$$

(with $\|\mathbf{z}\| \equiv \sqrt{\mathbf{z}^T\mathbf{z}}$) which could be interpreted as a nondimensional measure of the *stabilty* of the flow patterns.

- (Model courtesy of J. Zavala-Hidalgo and colleagues.)


30
28
26
24
22
20
-98
-96
-94
-92
-90
-88
-86
-84
-82
10
20
30
40
50
60
70
Velocity Standard Deviation (cm/s)


(b) HYCOM
30
28
26
24
22
20
-98
-96
-94
-92
-90
-88
-86
-84
-82
0
0.2
0.4
0.6
0.8
1
1.2
Ratio of Magnitude of Mean Flow to Standard Deviation

Statistics Example

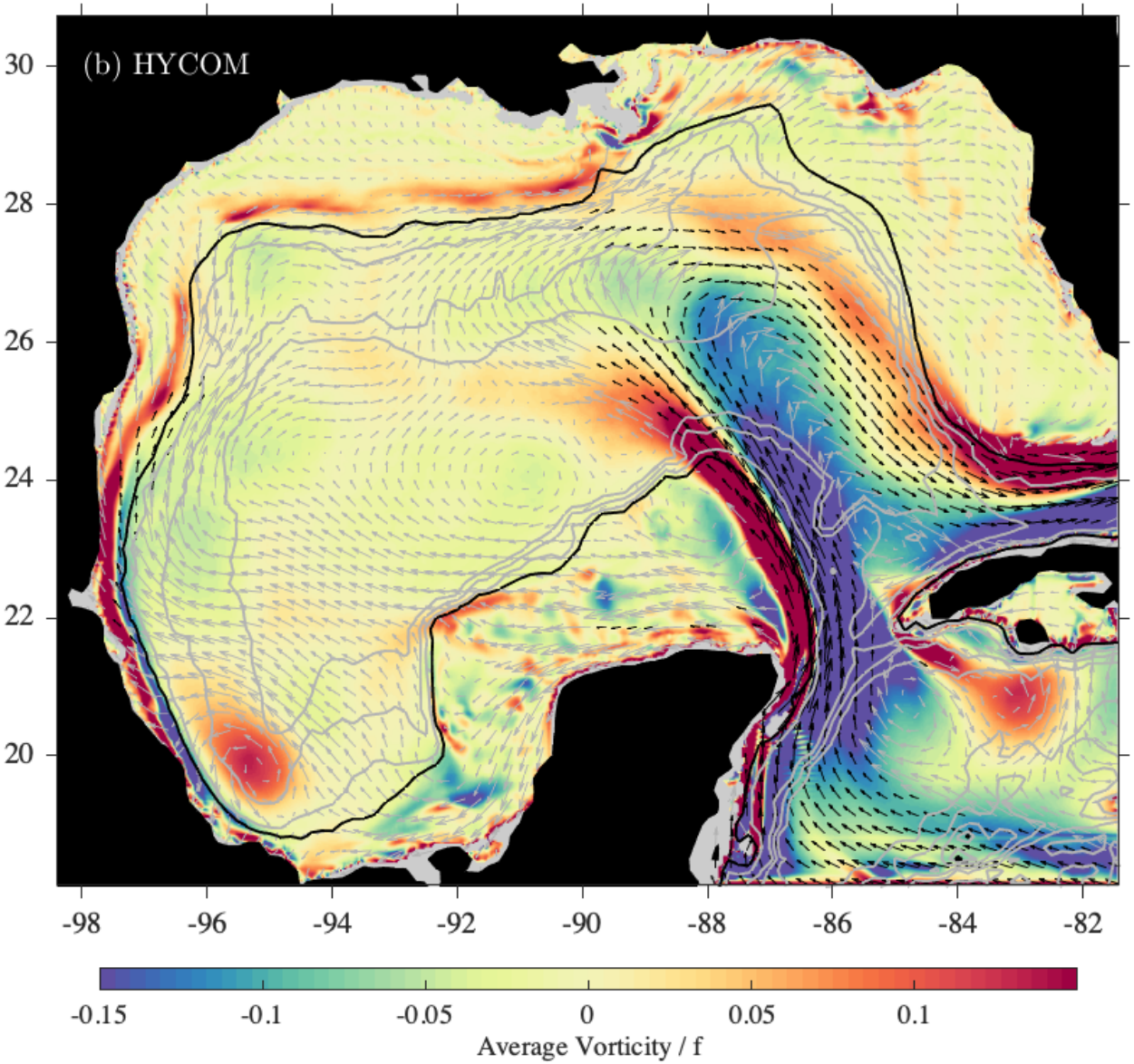
Next we will look at the first three moments of the vorticity

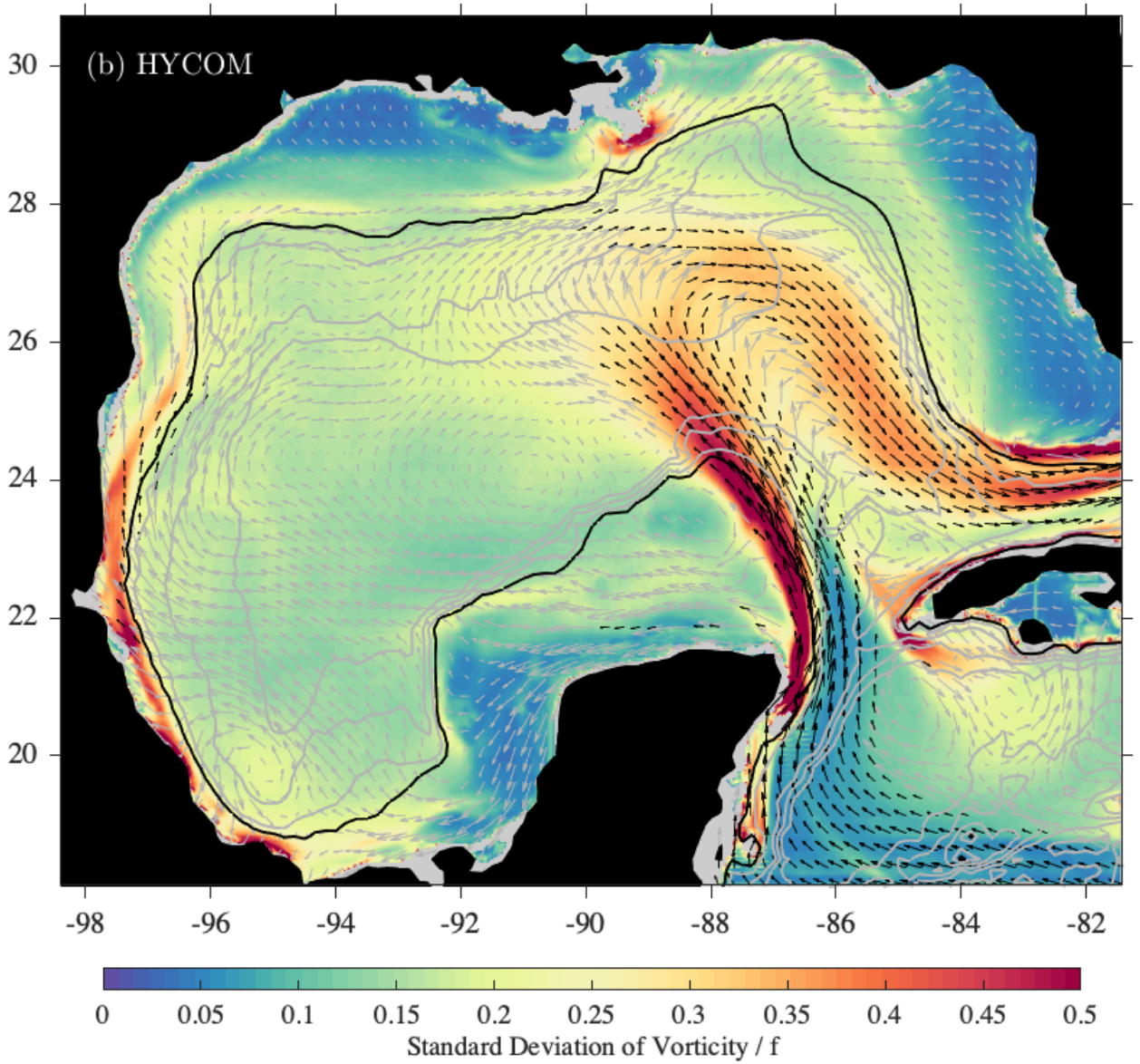
$$\zeta_n(x,y) \equiv \frac{\partial v_n}{\partial x} - \frac{\partial u_n}{\partial y}$$

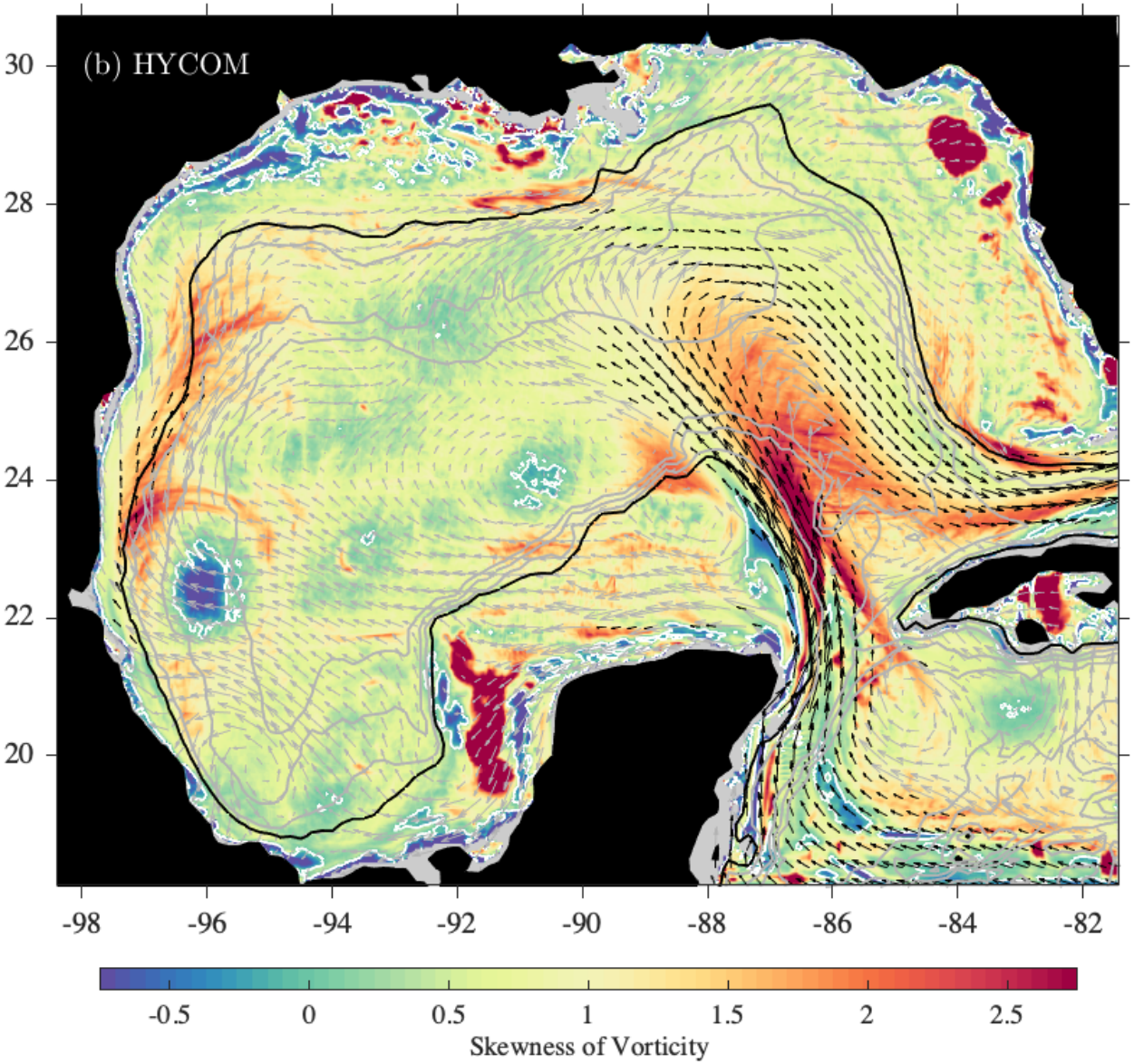
$$\bar{\zeta}(x,y) \equiv \frac{1}{N}\sum_{n=0}^{N-1} \zeta_n$$

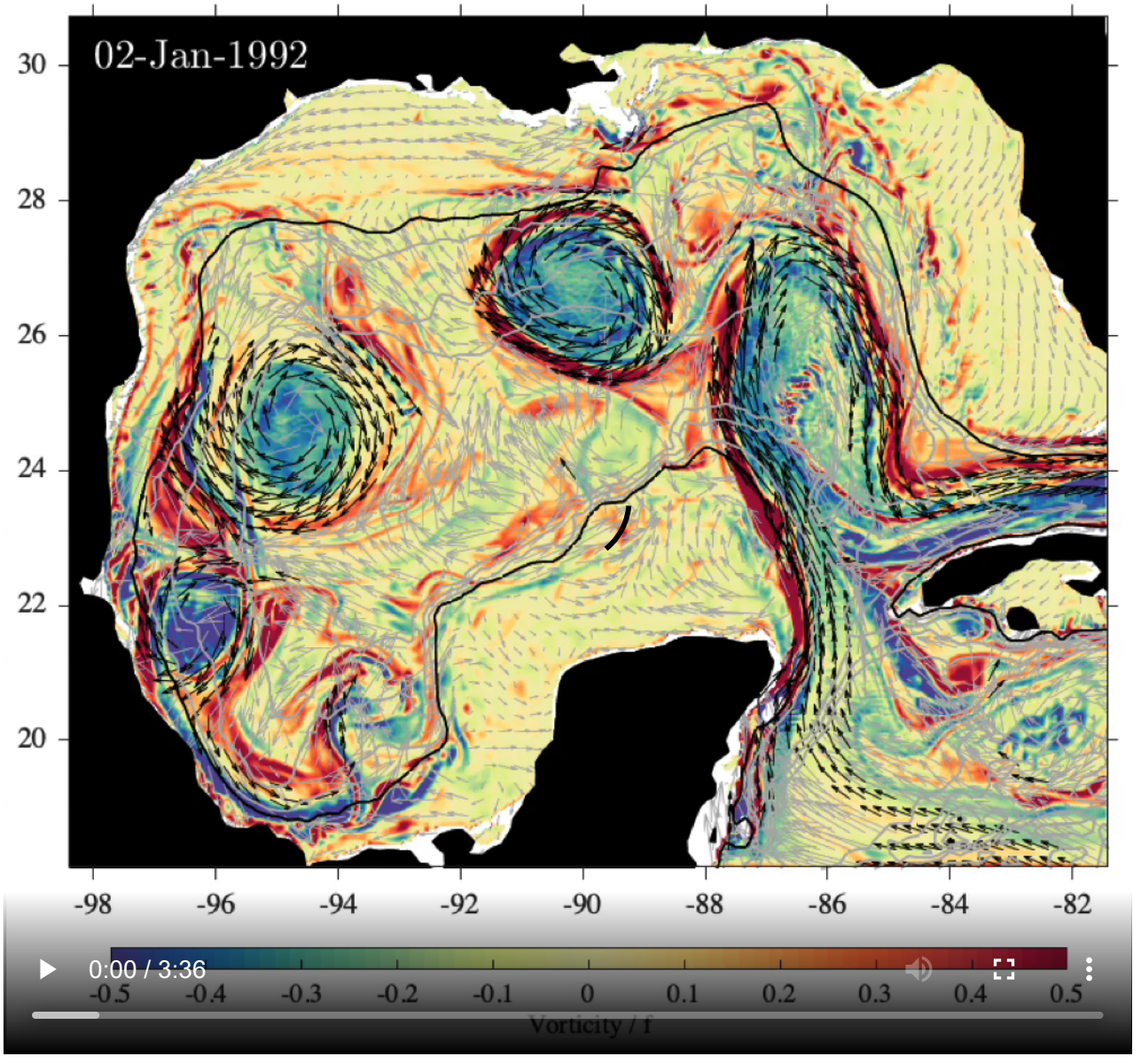
$$\sigma_\zeta^2(x,y) \equiv \frac{1}{N}\sum_{n=0}^{N-1} \left(\zeta_n - \bar{\zeta}\right)^2$$

$$\bar{\gamma}_\zeta(x,y) \equiv \frac{1}{\sigma_\zeta^3}\frac{1}{N}\sum_{n=0}^{N-1} \left(\zeta_n - \bar{\zeta}\right)^3.$$


(b) HYCOM
30
28
26
24
22
20
-98
-96
-94
-92
-90
-88
-86
-84
-82
-0.15
-0.1
-0.05
0
0.05
0.1
Average Vorticity / f


(b) HYCOM
30
28
26
24
22
20
-98
-96
-94
-92
-90
-88
-86
-84
-82
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Standard Deviation of Vorticity / f


(b) HYCOM
30
28
26
24
22
20
-98
-96
-94
-92
-90
-88
-86
-84
-82
-0.5
0
0.5
1
1.5
2
2.5
Skewness of Vorticity


02-Jan-1992
30
28
26
24
22
20
-98
-96
-94
-92
-90
-88
-86
-84
-82
0:00 / 3:36
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
Vorticity / f

Takeaway Messages

The important message here is that time-domain statistics, while being an important tool, don't capture all the complex structure of the time-evolving turublent ocean.

The time-domain statistics "flatten" the variability, the way a shadow flattens a three-dimensional object. They provide us with compact summaries but at the expense of compressing the rich structure.

Higher-order statistics—the skewness and kurtosis—can *sometimes* reveal features that are hidden by the lower-order statistics.

It's often a great idea to make an animation!

Homework

1. Compute and plot the histogram of your data. In Matlab, you can do this using Matlab's `hist` or `histogram` functions. (For bivariate data, do this and the next step separately for both components.)
2. Compute the sample mean and variance.
3. Experiment with filtering your data. In Matlab, this can be done using `vfilt`. Plot the data, the filtered version, and the residual (original minus filtered) for a few choices of filter length. Are there any choices that seems to be suitable for your data?
4. Re-do the steps 1&2 involving the time-domain statistics, but using firstly the smoothed, and secondly the residual, versions of your data. How do the statistics change dependent upon the lowpass or highpass filtering? How do you interpret this?

A Gallery of Visualizations

Overview

In this lecture we'll look at some examples of data visualizations and simple analyses from several papers.

We will give names to different plot types as way to understand generic approaches that can be applied in different situations.

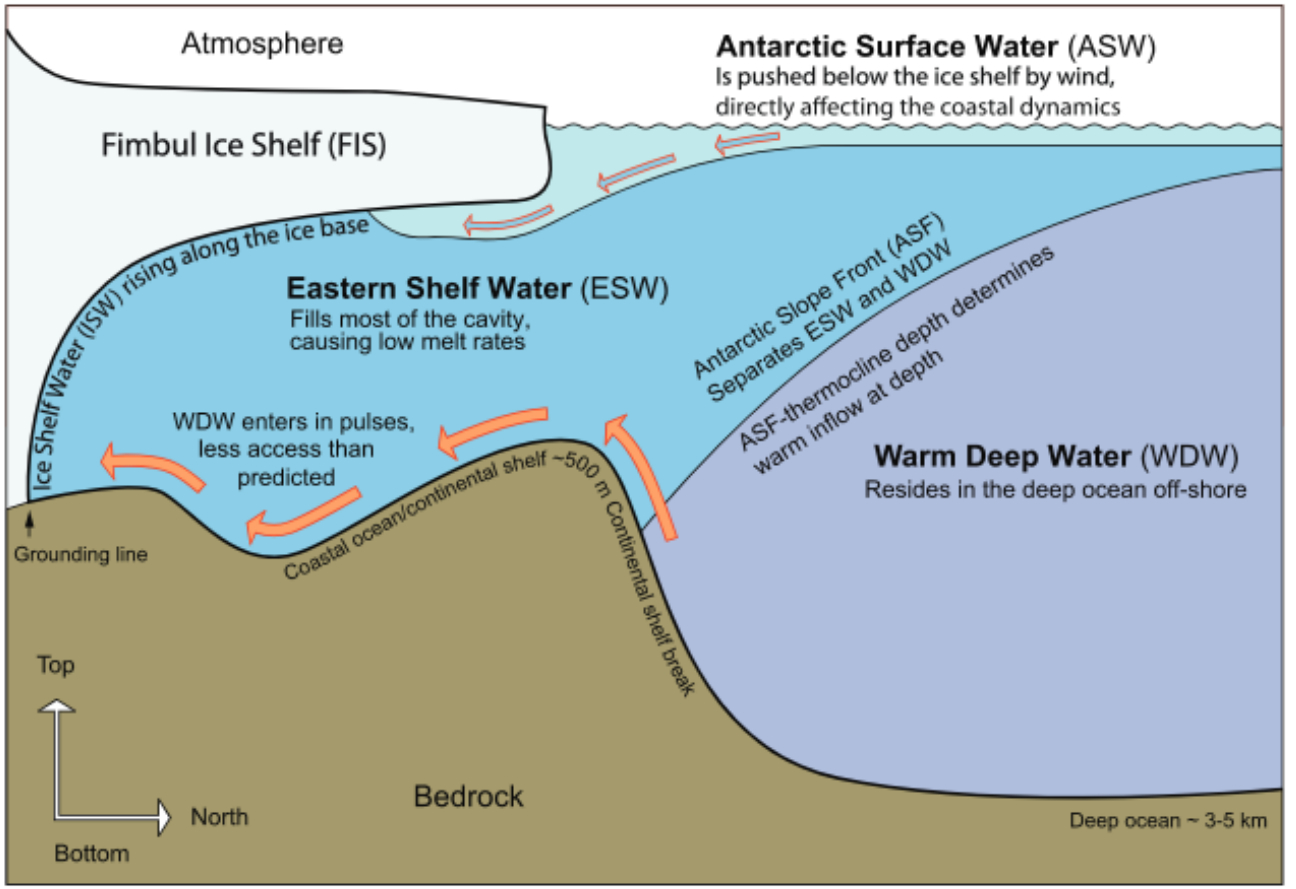
Try to appreciate them from a great distance, understanding how they are presenting information rather than the details.

Imagine how you might use each of these to examine your data, and makes notes of which plot types would be most likely to be useful.

Some examples from

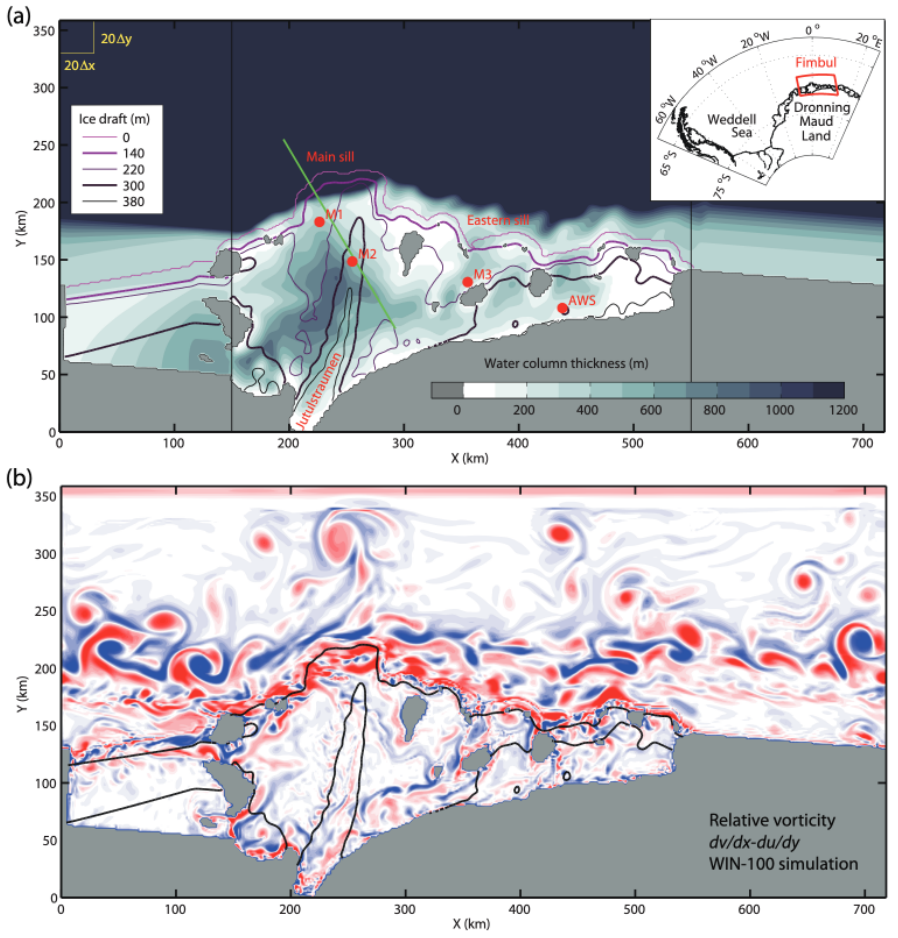
Hattermann, T., L. H. Smedsrud, O. A. Nøst, J. M. Lilly, and B. K. Galton-Fenzi (2014). Eddy-resolving simulations of the Fimbul Ice Shelf cavity circulation: Basal melting and exchange with open ocean. {link}

The Schematic



A visual representation of your hypothesis of how the system works.

The Orientation Plot + The Snapshot

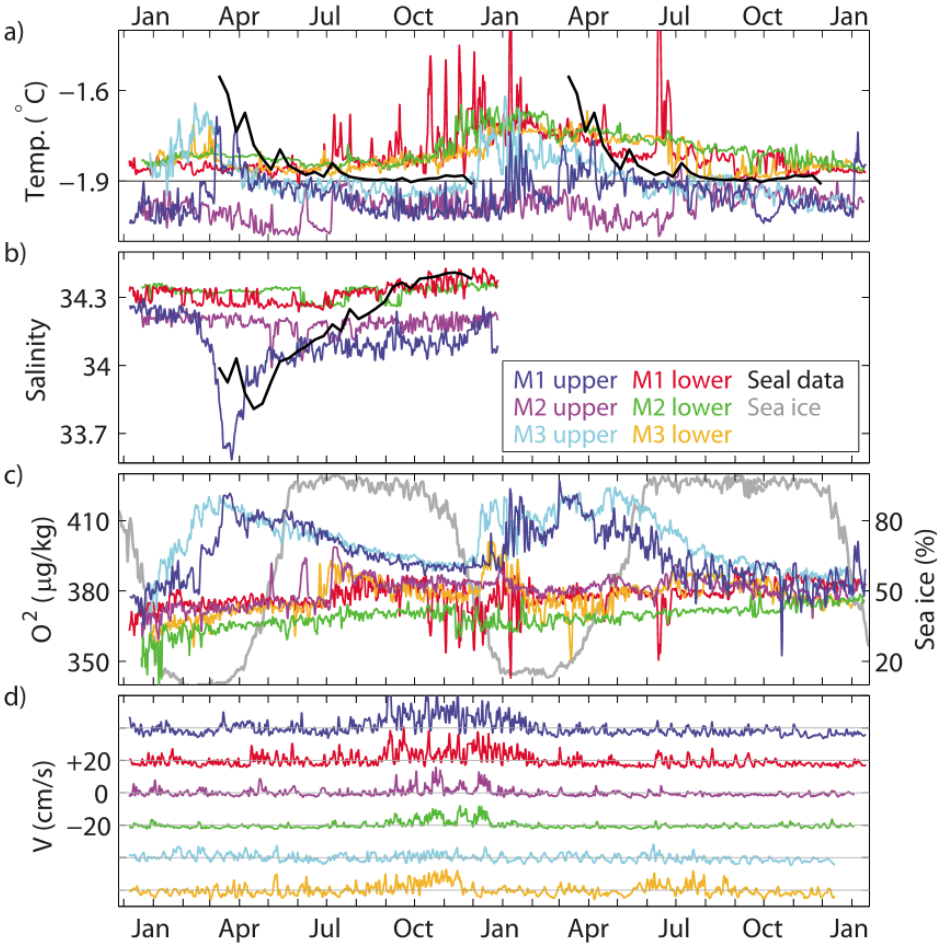


This combination of two plots is more meaningful than either alone.

An example from

Hattermann, T., O. A. Nøst, J. M. Lilly, and L. H. Smedsrud (2012). Two years of oceanic observations below the Fimbul Ice Shelf, Antarctica. {link}

The Offset Line Plot

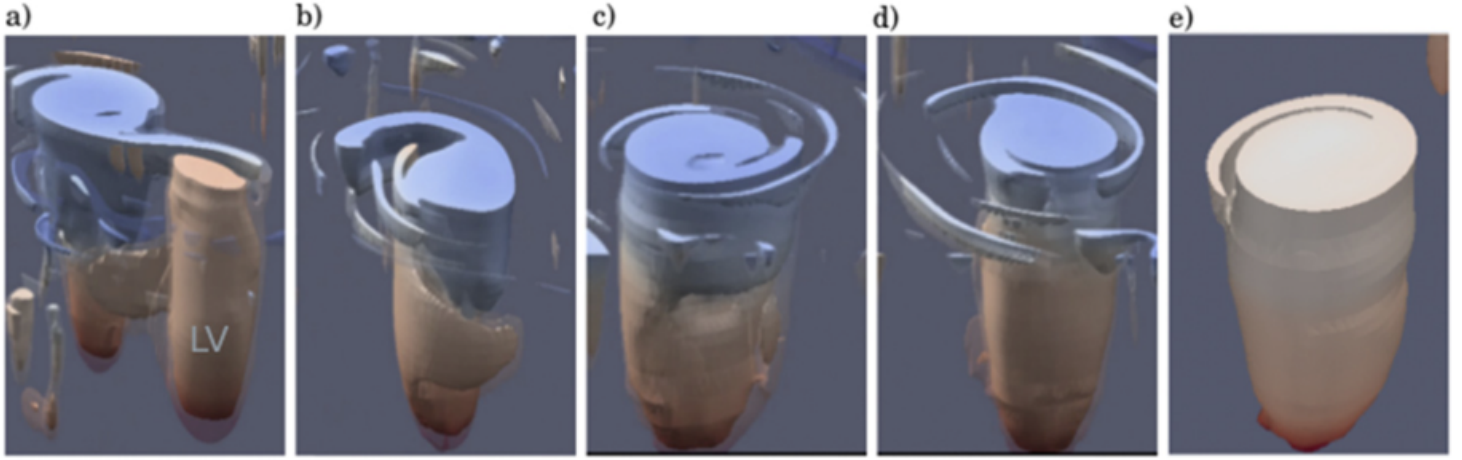


Note (i) stacked axes and (ii) the very informative use of color.

An example from

Trodahl, M., P. E. Isachsen, J. Nilsson, J. M. Lilly, and N. M. Kristensen (2020). The regeneration of the Lofoten Vortex through vertical alignment. {link}

The Series of Snapshots



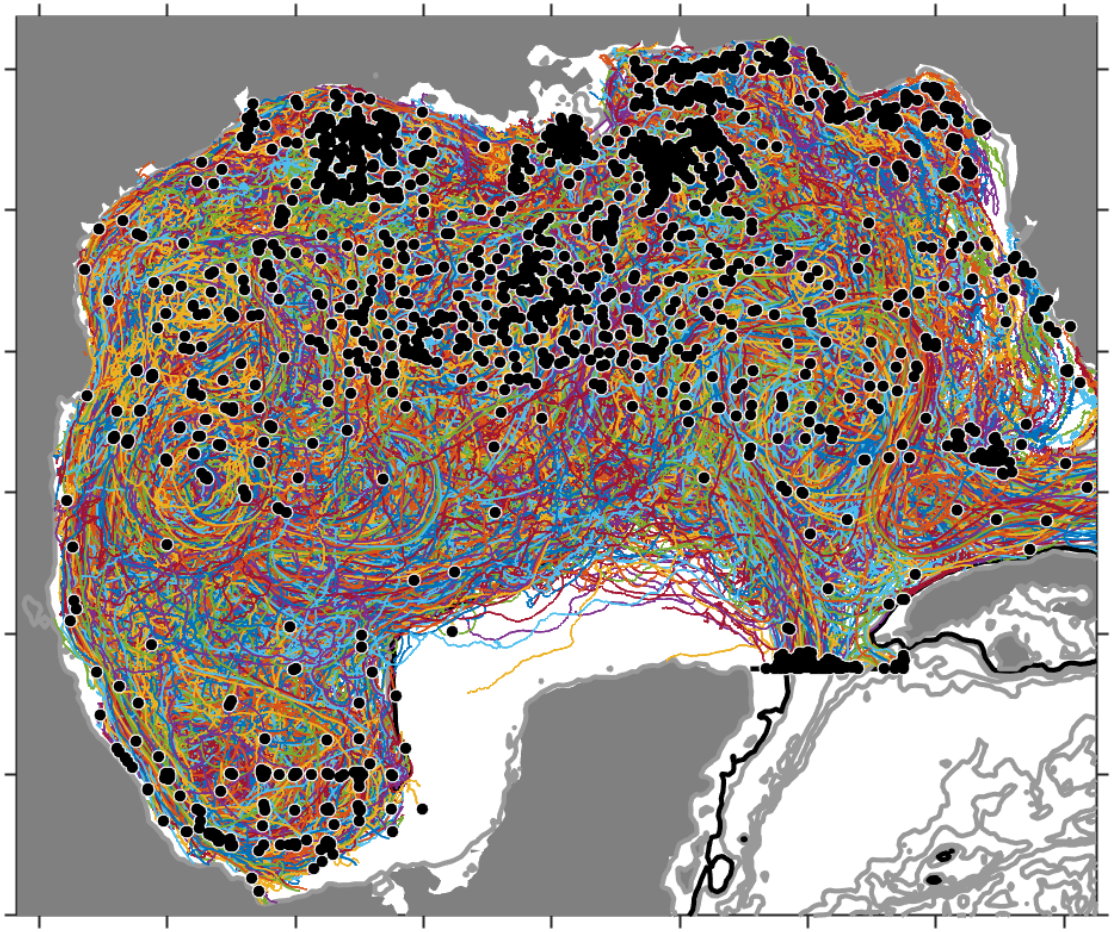
The eye fills in the gaps, giving the impression of a movie.

There are five (!) dimensions visualized in this plot: x, y, z, time, and density (color).

Some examples from

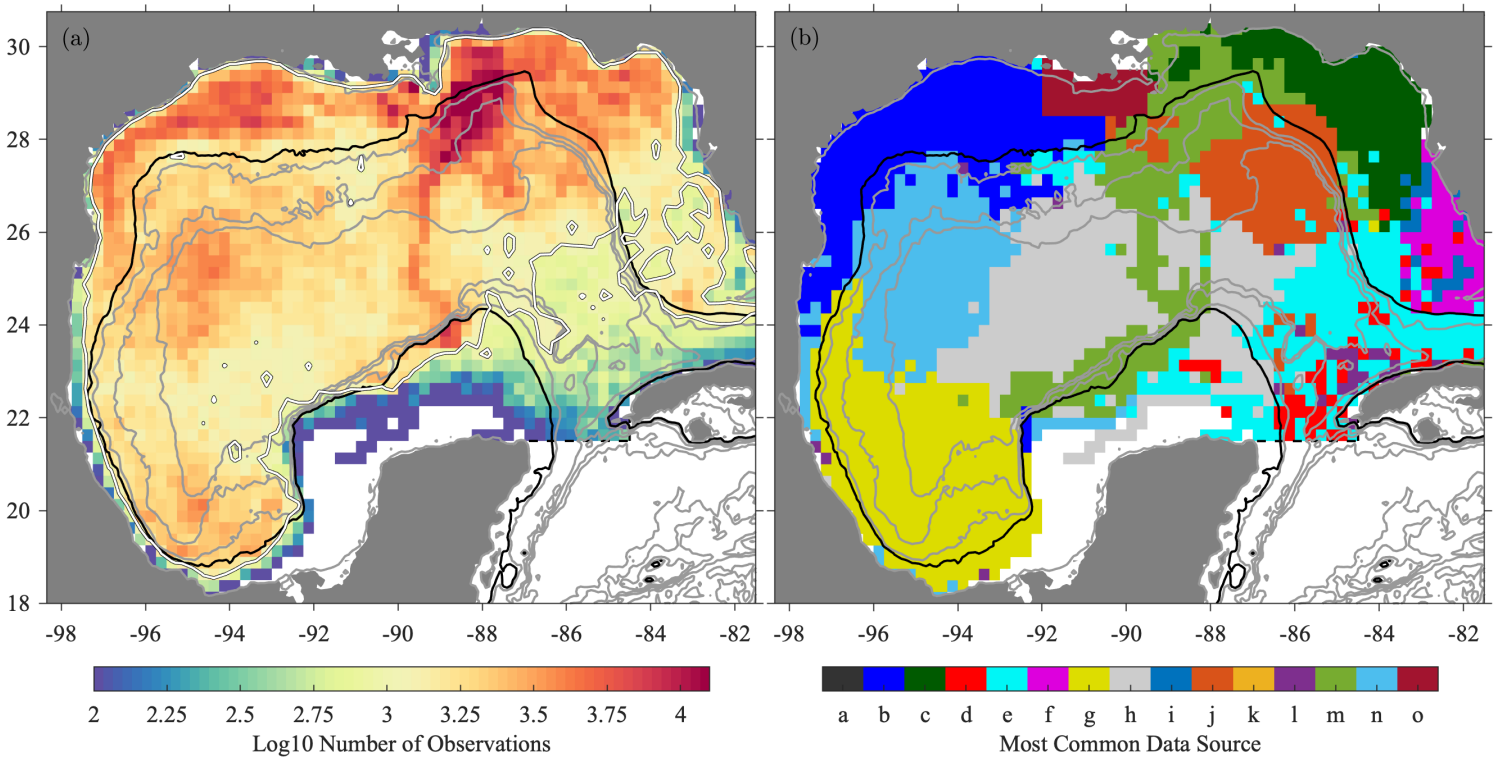
Lilly, J. M. and P. Pérez-Brunius (2021b). Extracting statistically significant eddy signals from large Lagrangian datasets using wavelet ridge analysis, with application to the Gulf of Mexico. {link}

The Spaghetti Plot



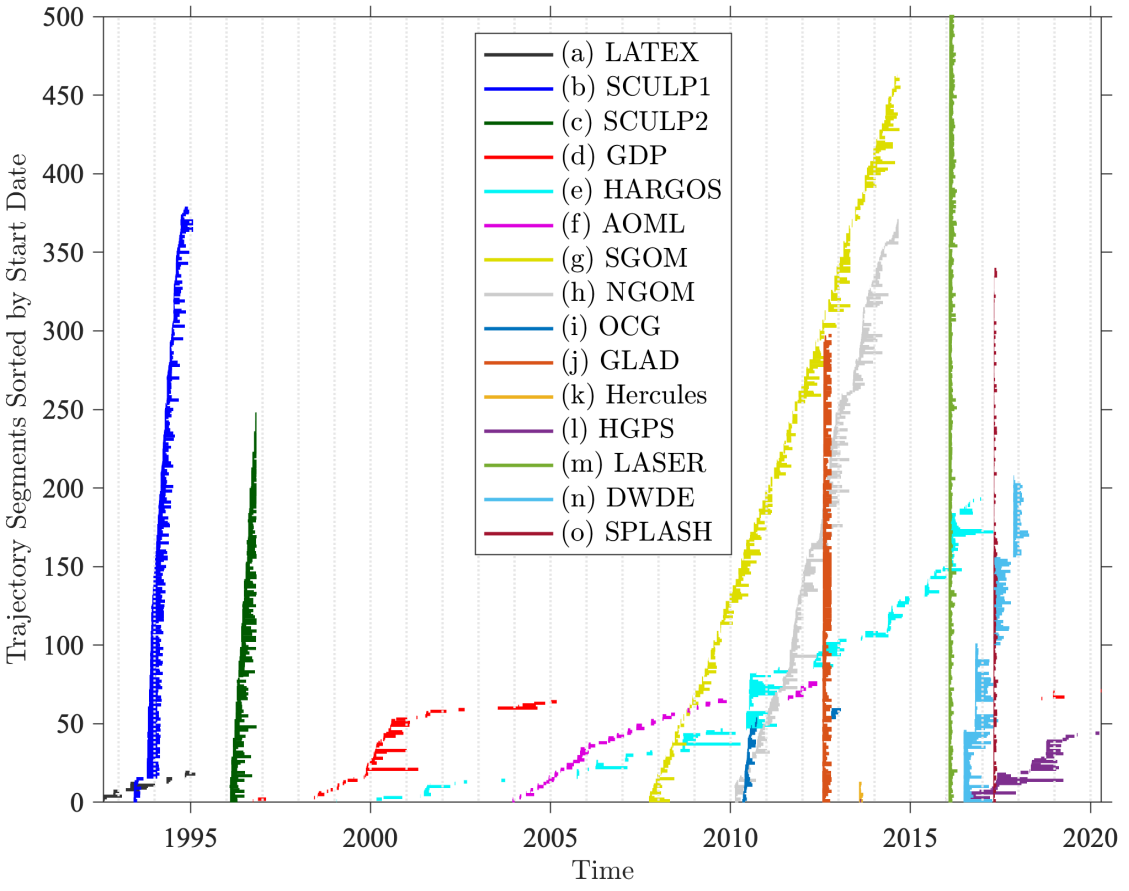
A plot of the trajectories along which measurements are taken. Note use of symbols to denote trajectory start point.

2D Histogram and Classification Map



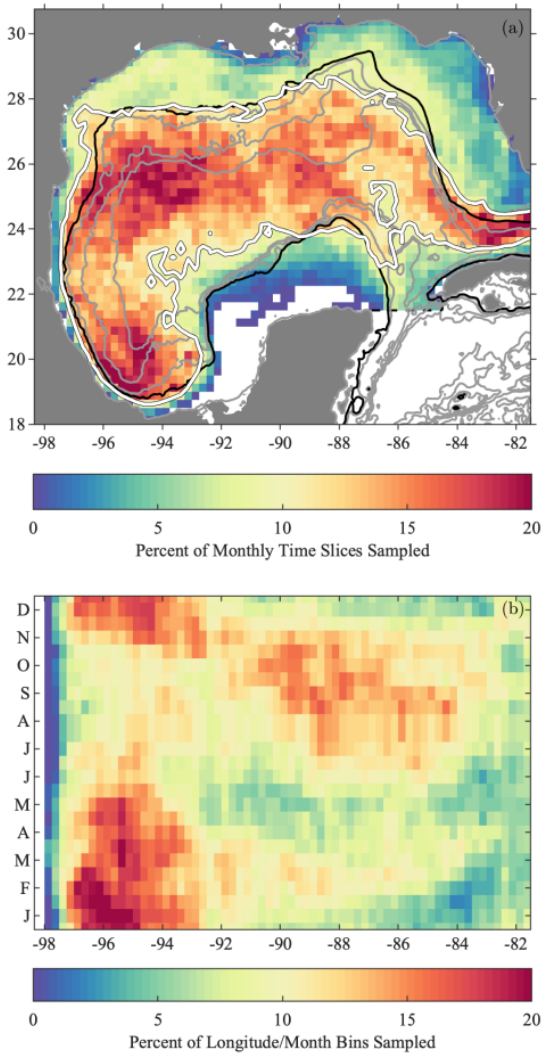
Log10 number of observations in bins from 15 different experiments (left). Most common experimental source in each bin (right).

Creative Line Plot



Information in (a) line color (b) line length (c) line x-start point and (d) line y-start point.

Dovetailing 2D Histograms

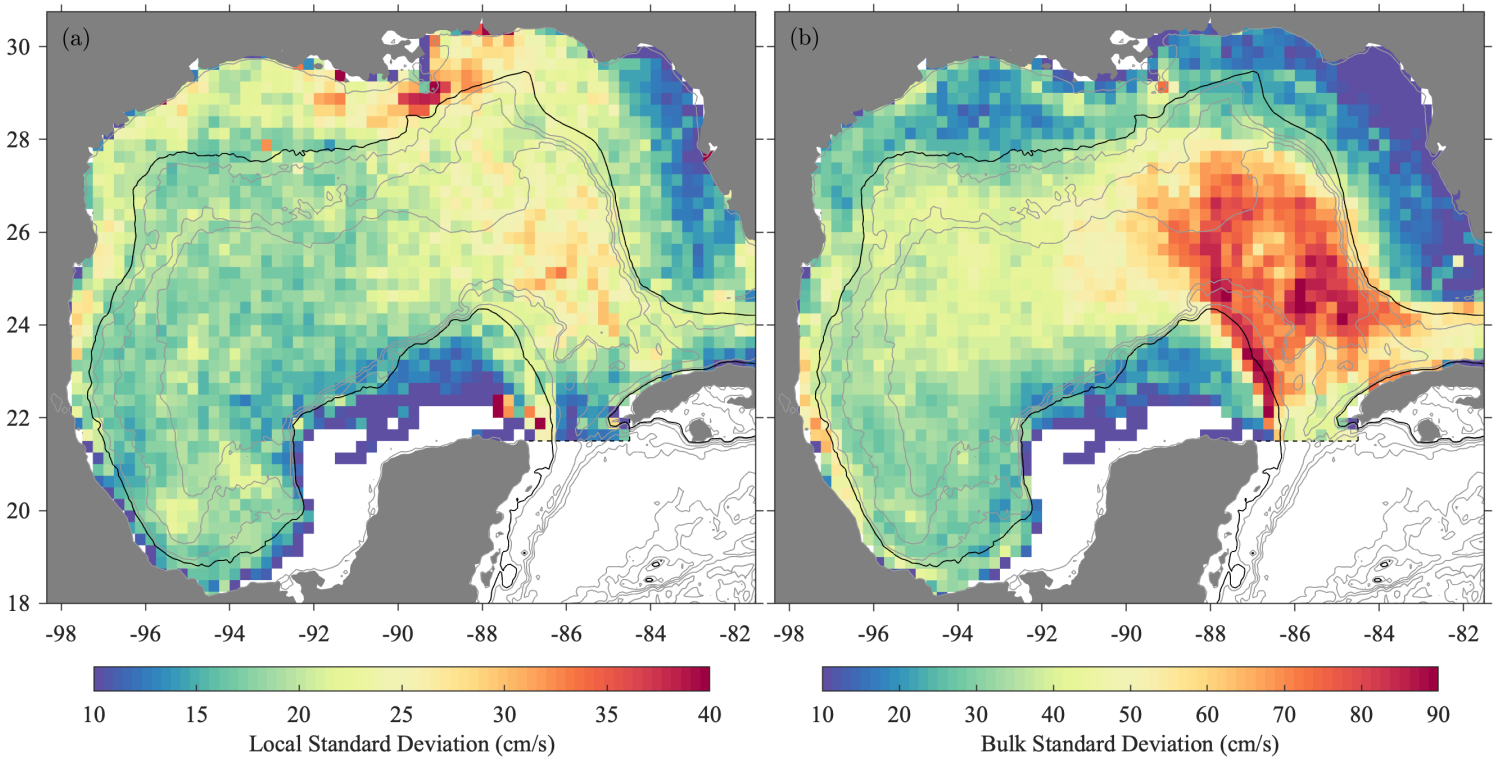


Top: Histogram of month-long time periods sampled in each latitude/longitude bin.

Bottom: The same information, but now in longitude/month bins.

Note the x-axes are the same, i.e. we are looking at 3D distribution in two 2D slices.

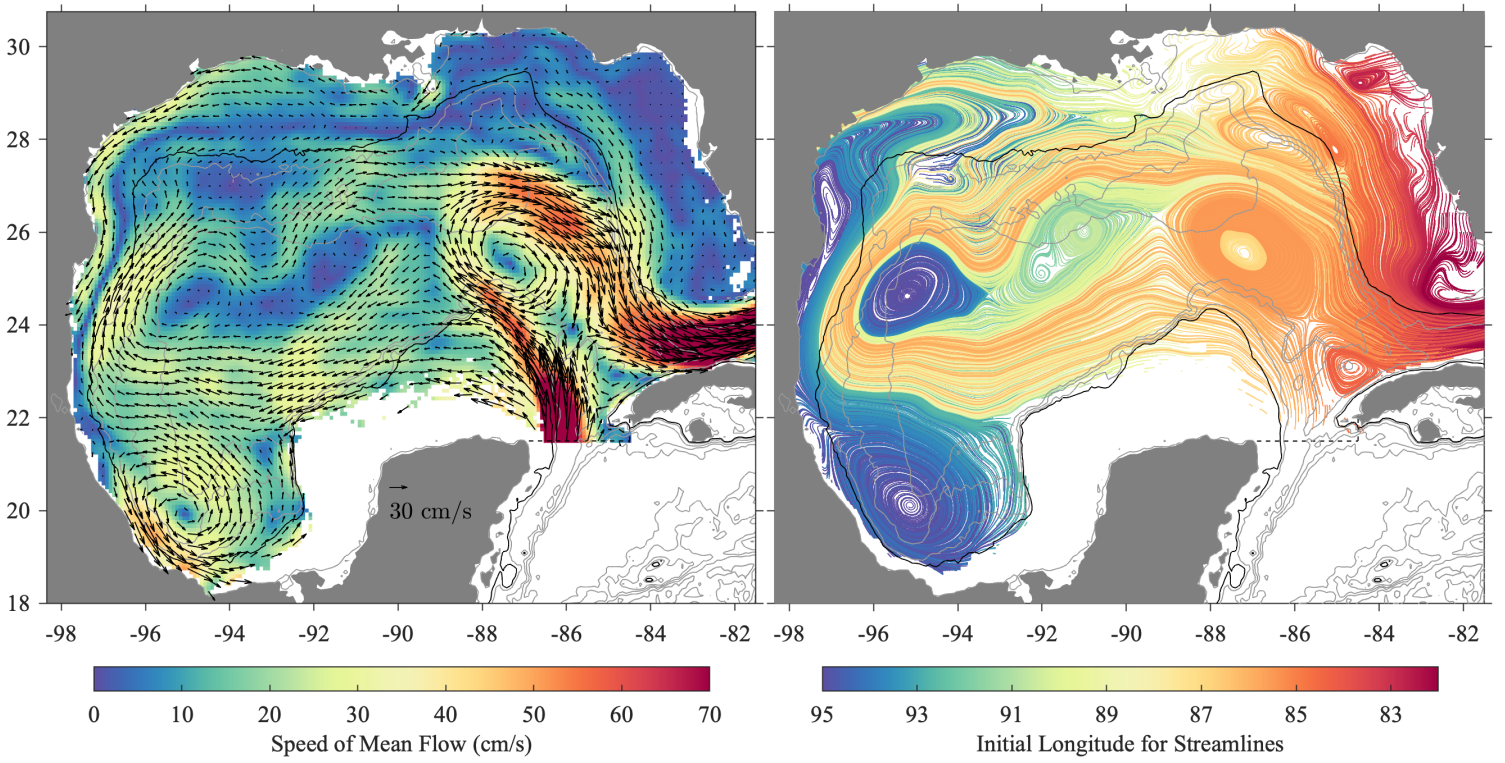
Downscaling Dataset Resolution



Right: Variability of the average in each bin about the time average.
Left: Variability within each bin to the time-dependent bin average.

The variability on the left is the variability that we lose when we grid the data. We can keep track of this lost variability explicitly.

Shaded Vector Plot and Streamline Plot



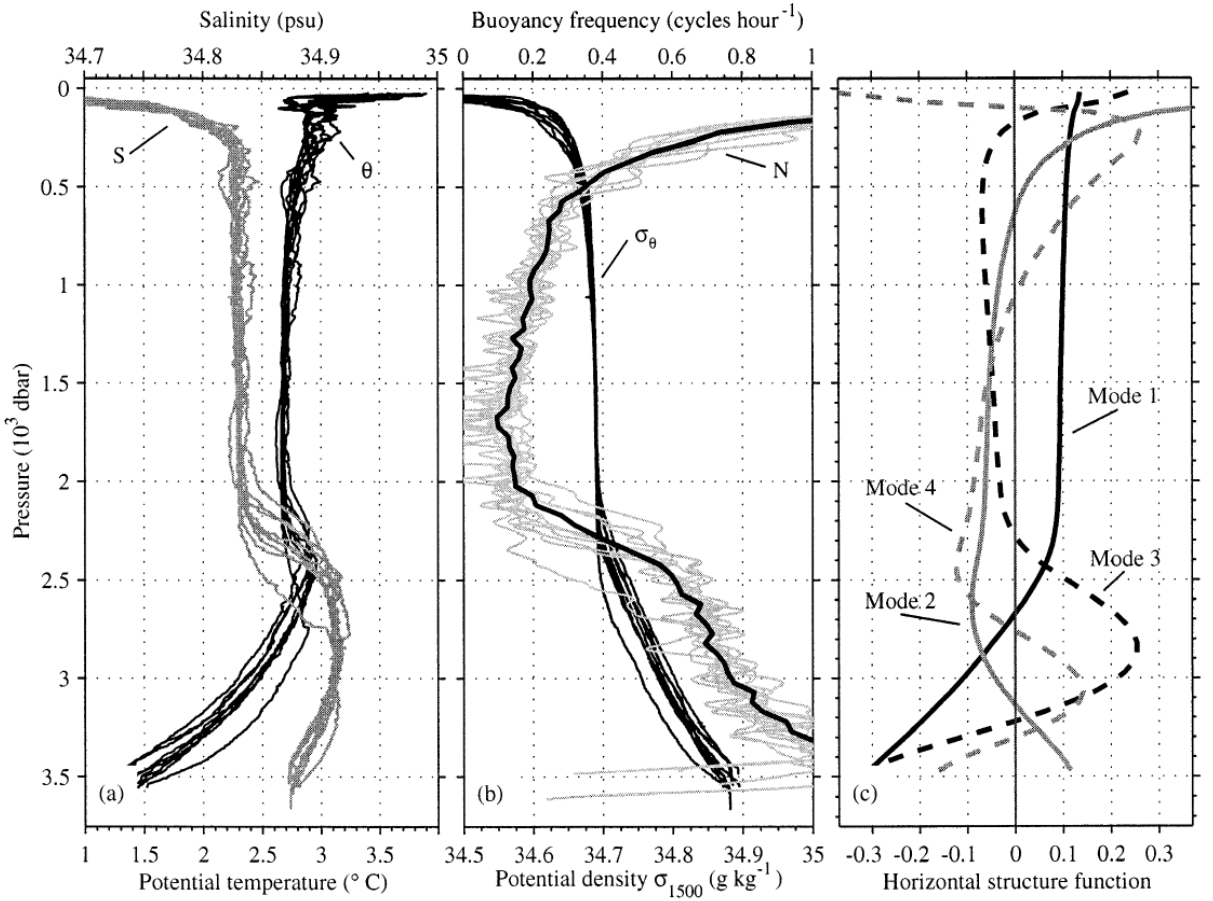
Left: mean flow field displayed as arrows, superposed on color shading showing, in this case, the mean flow magnitude.

Right: the same mean flow field visualized as artificial particle trajectories, a.k.a. a streamline plot. Color shows initial longitude.

Some examples from

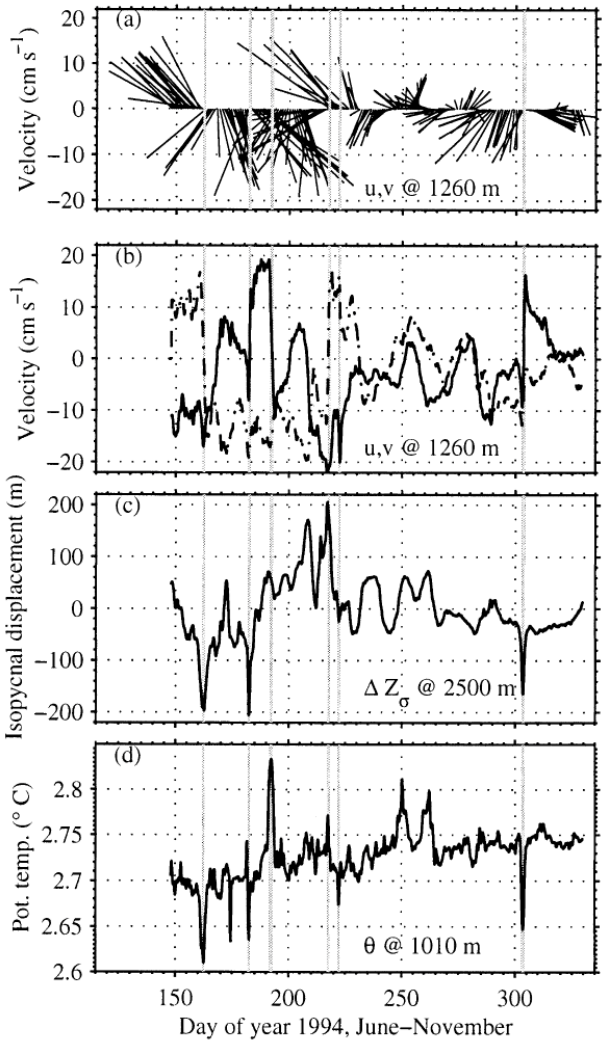
Lilly, J. M. and P. B. Rhines (2002). Coherent eddies in the Labrador Sea observed from a mooring. {link}

Simple Line Plot



Simple line plots can be very useful, especially when labelled. Note the “Mean+Variability” plot in the middle.

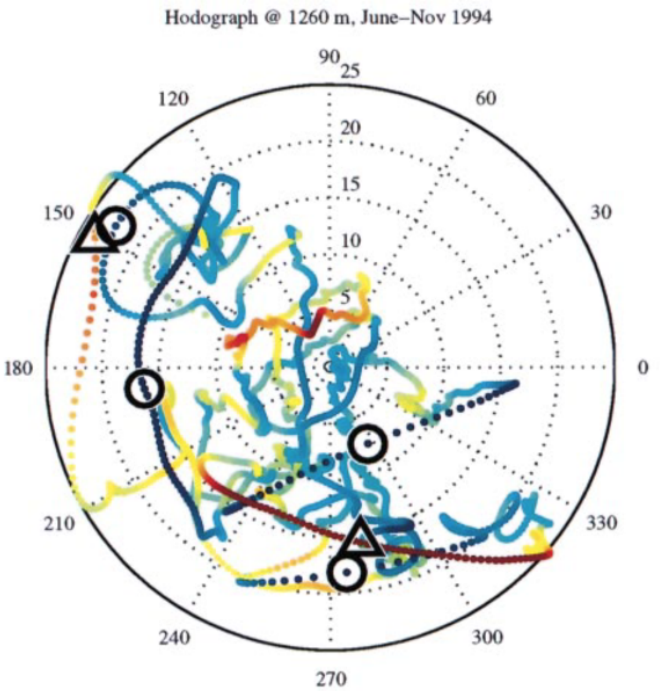
The Event Plot

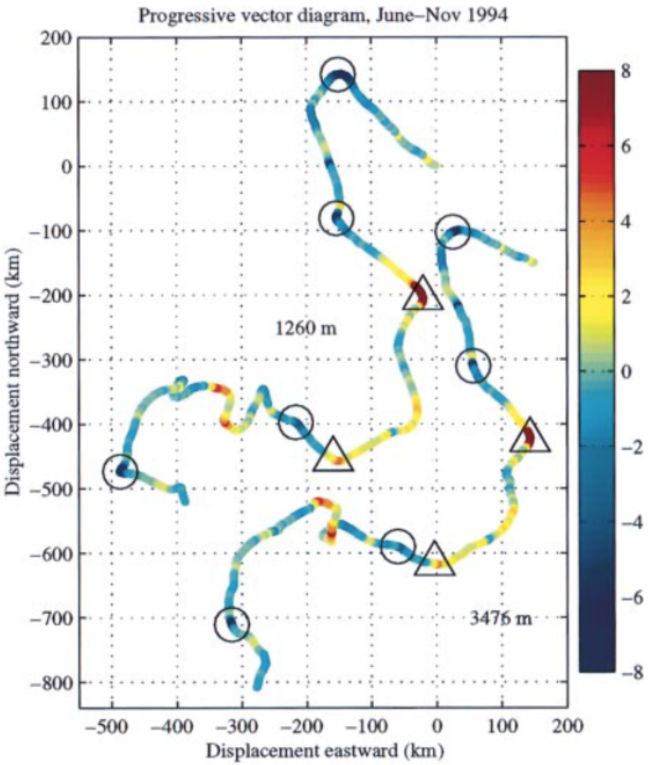


Multiple fields are shown with apparent events indicated, in this case, by a line.

The top panel is “Stickvector Plot” of velocity.

Hodographs and Progressive Vectors

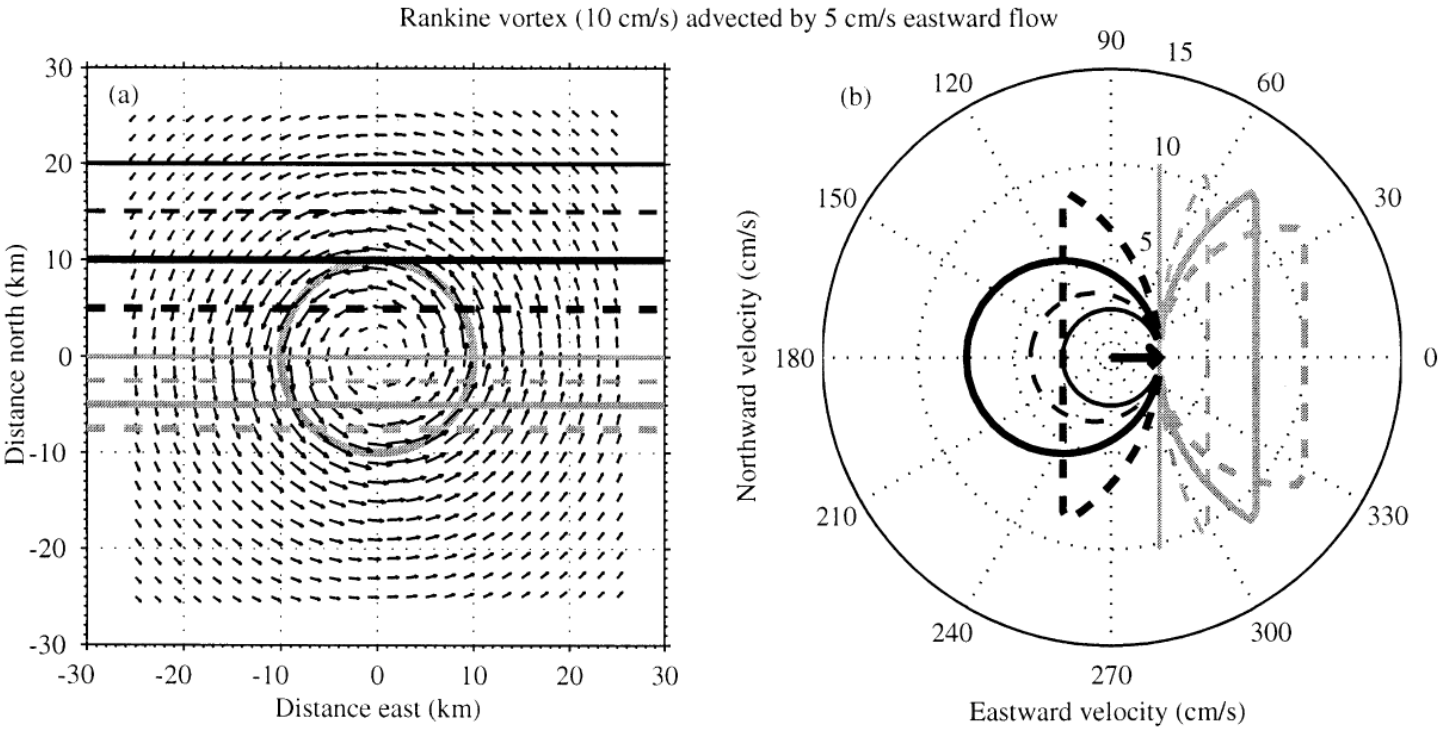




Left: A hodograph, the curve traced out by the velocity vector.
Right: The progressive vector diagram, the integral of the velocity.

These are examples of “3D Scatter Plots” with color showing a third quantity.

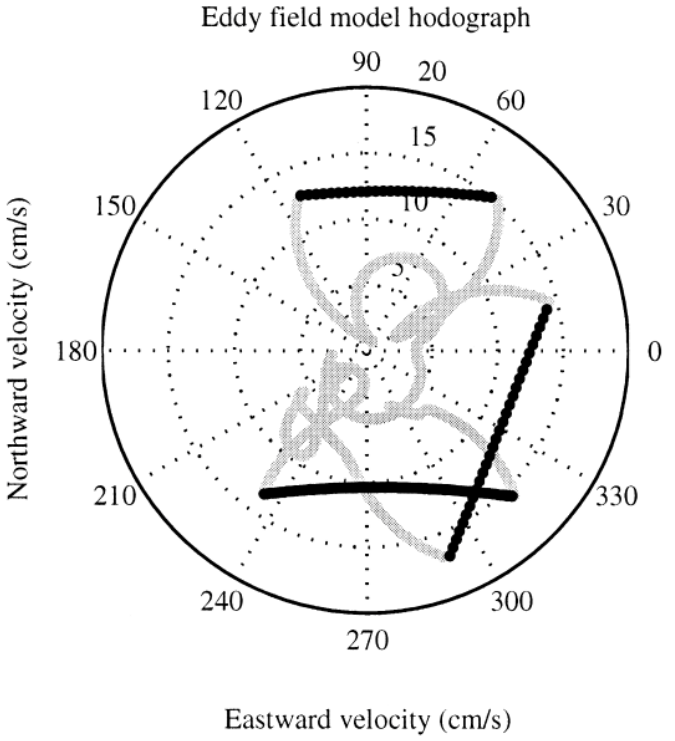
Kinematic Model

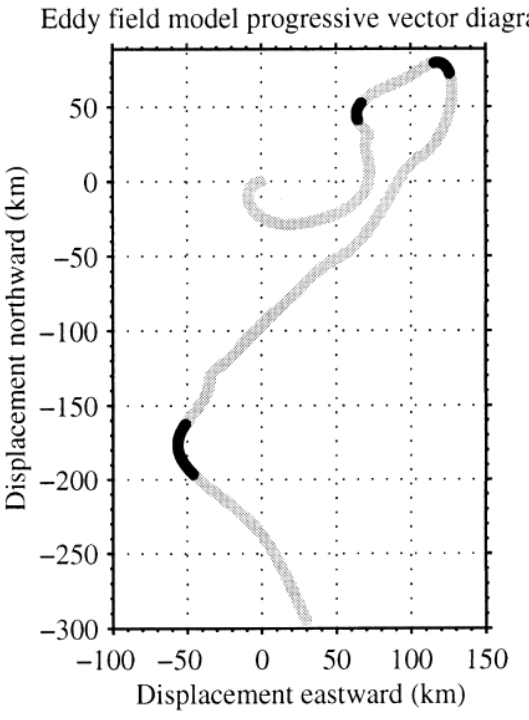


An often powerful approach is to come up with a simple conceptual model (that is, a hypothesis) for what you think is happening.

In this case, we look at what happens when a Gaussian eddy is advected past a mooring.

Synthetic Observations





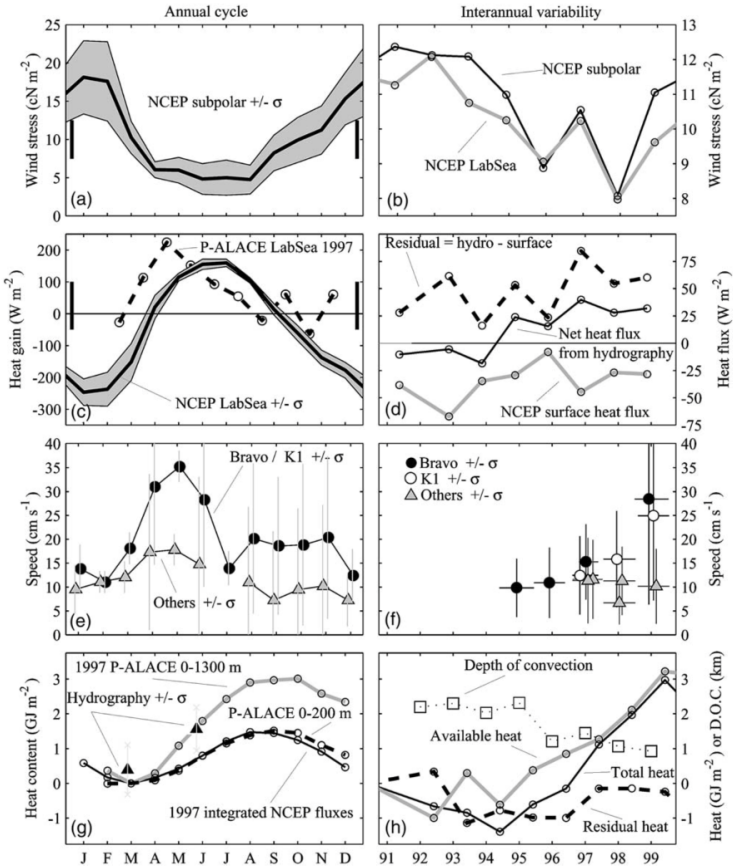
Then we may try to recreate quasi-realistic observations. This can be a powerful way to scrutinize proposed hypotheses.

Here, we have a hodograph and progressive vector diagram generated by several eddies advecting each other. Look familiar?

Some examples from

Lilly, J. M., P. B. Rhines, F. Schott, K. Lavender, J. Lazier, U. Send, and E. D'Asaro (2003). Observations of the Labrador Sea eddy field. {link}

The Kitchen Sink Plot

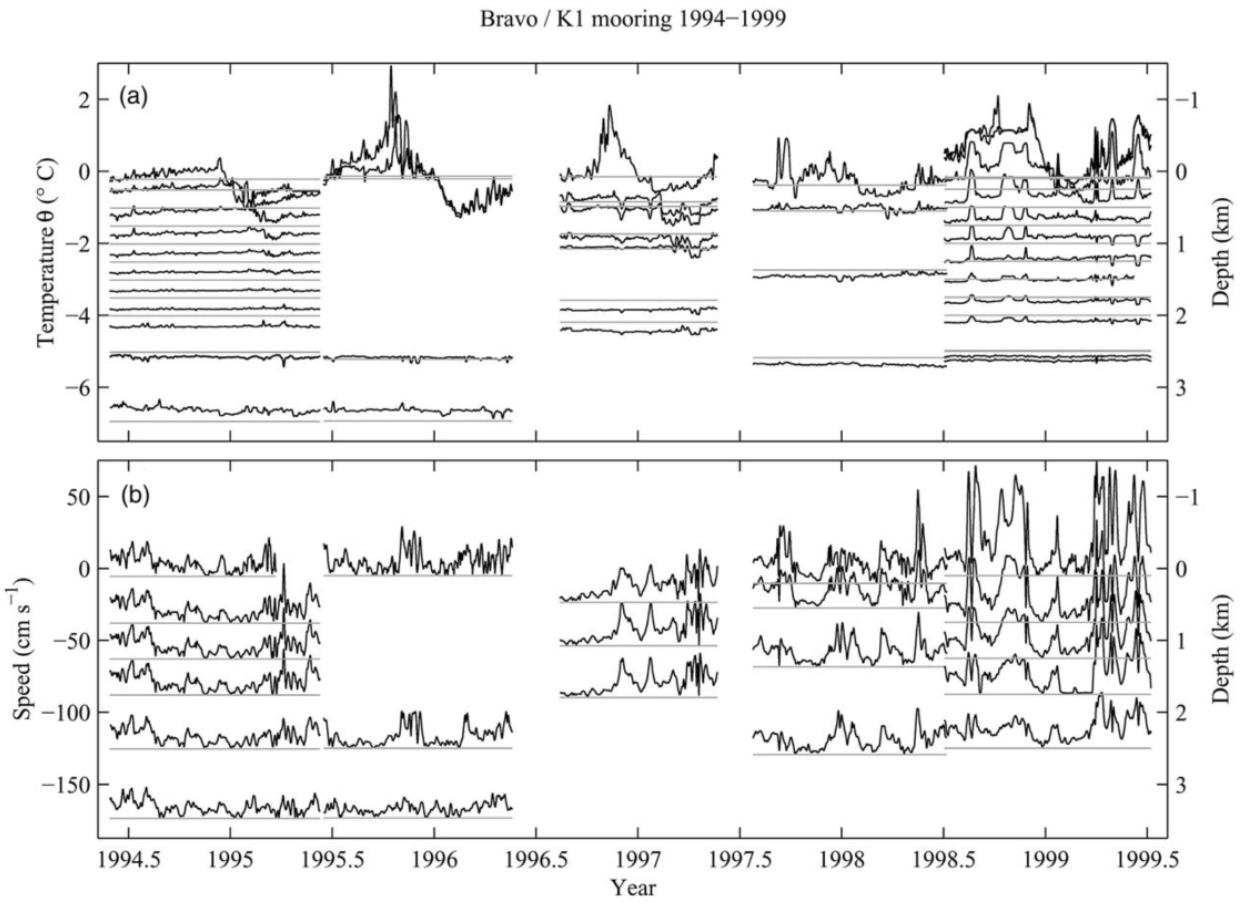


If you have a lot of datasets or fields that are interrelated, one approach is to try to show them all from some common perspective.

Here, we're looking at the annual cycle (left) and interannual variability (right) in various fields.

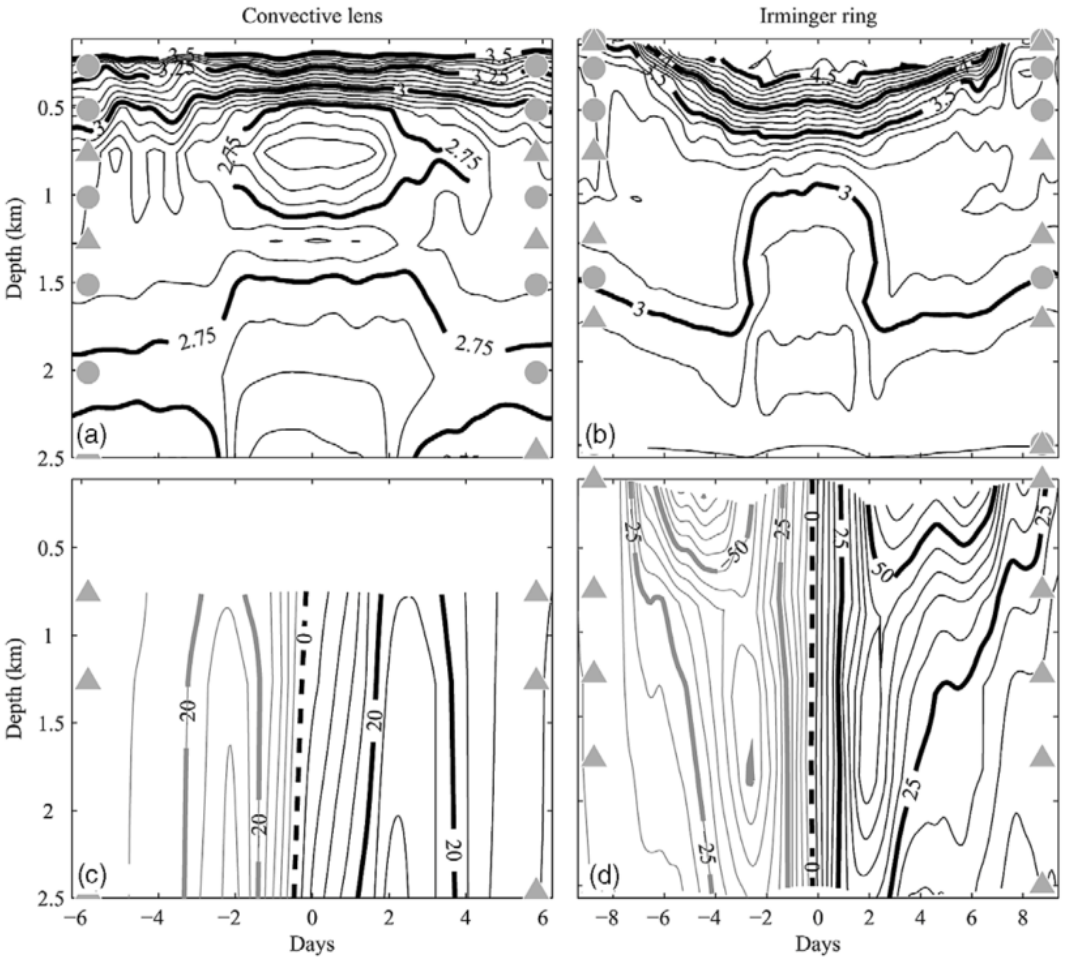
Small detail, note the flipped y-axis locations, the bottom two of which are shared between left and right.

The Offset Line Plot



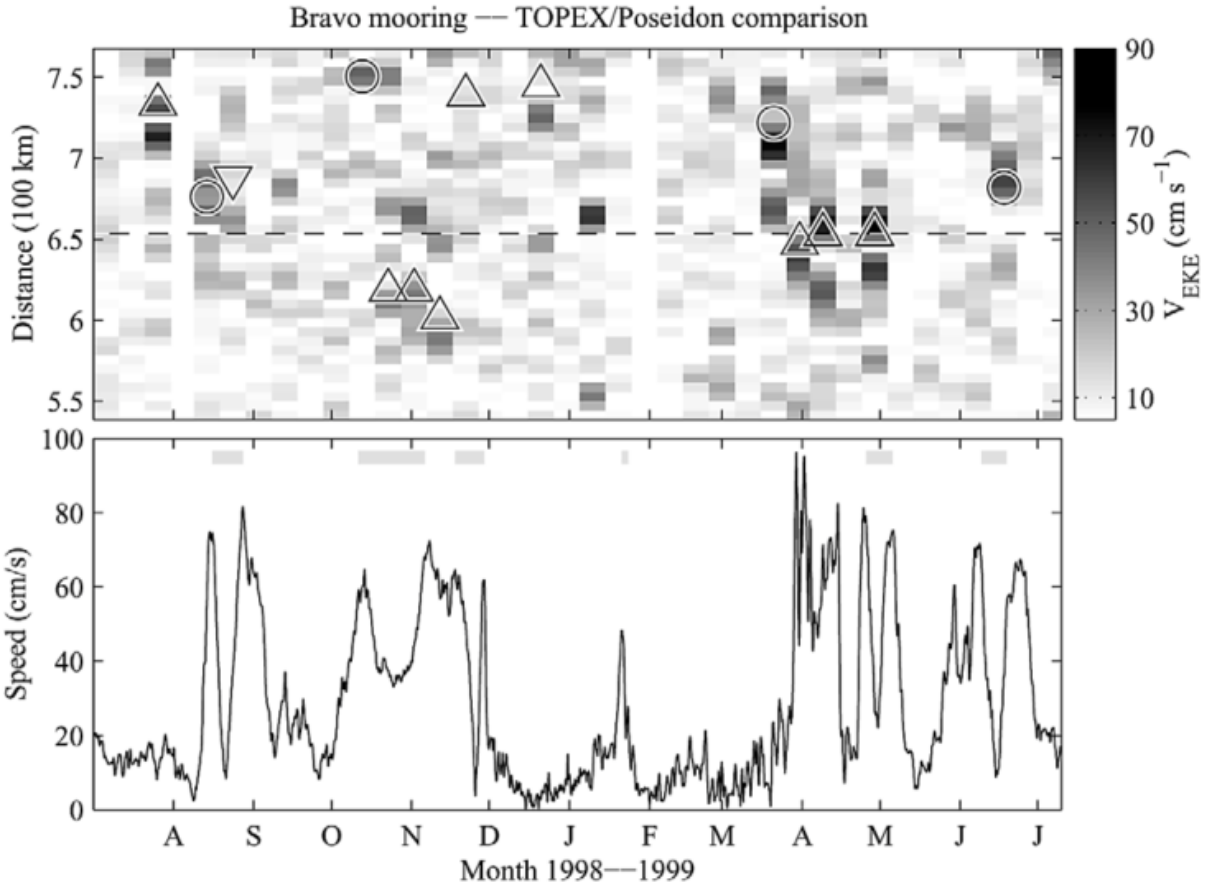
Lines show the measured variables, as well as the depth and time over which the measurements were taken.

The Representative Event Plot



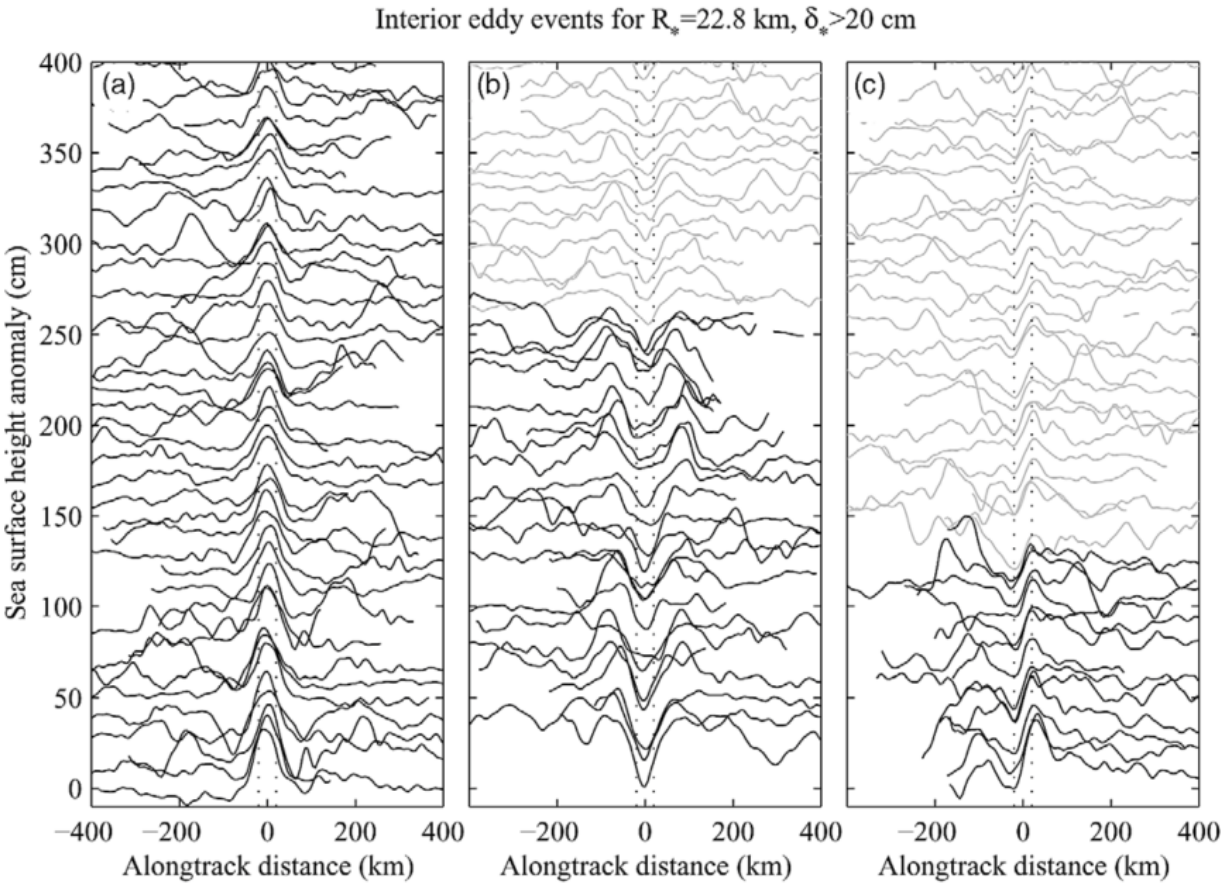
When analyzing multiple events, showing examples is useful.

Colocated Event Observations



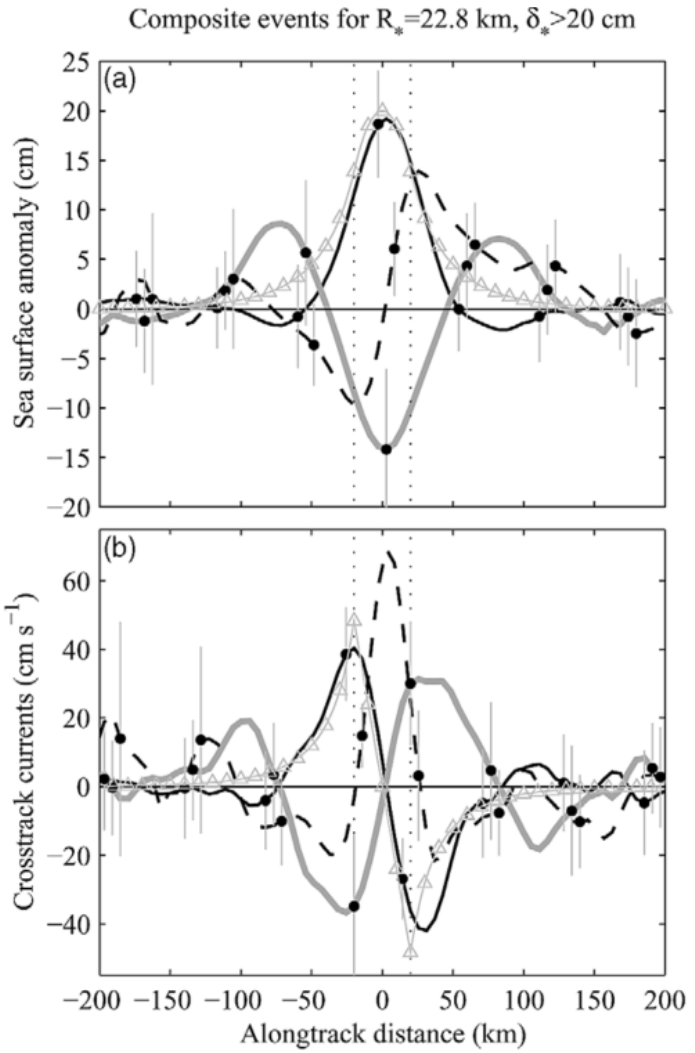
When different platforms observe similar events, you can try to find instances where both observe the same event at the same time.

The Waterfall Plot



A waterfall plot, or dense offset line plot, is useful when we're trying to draw attention to common features, for example.

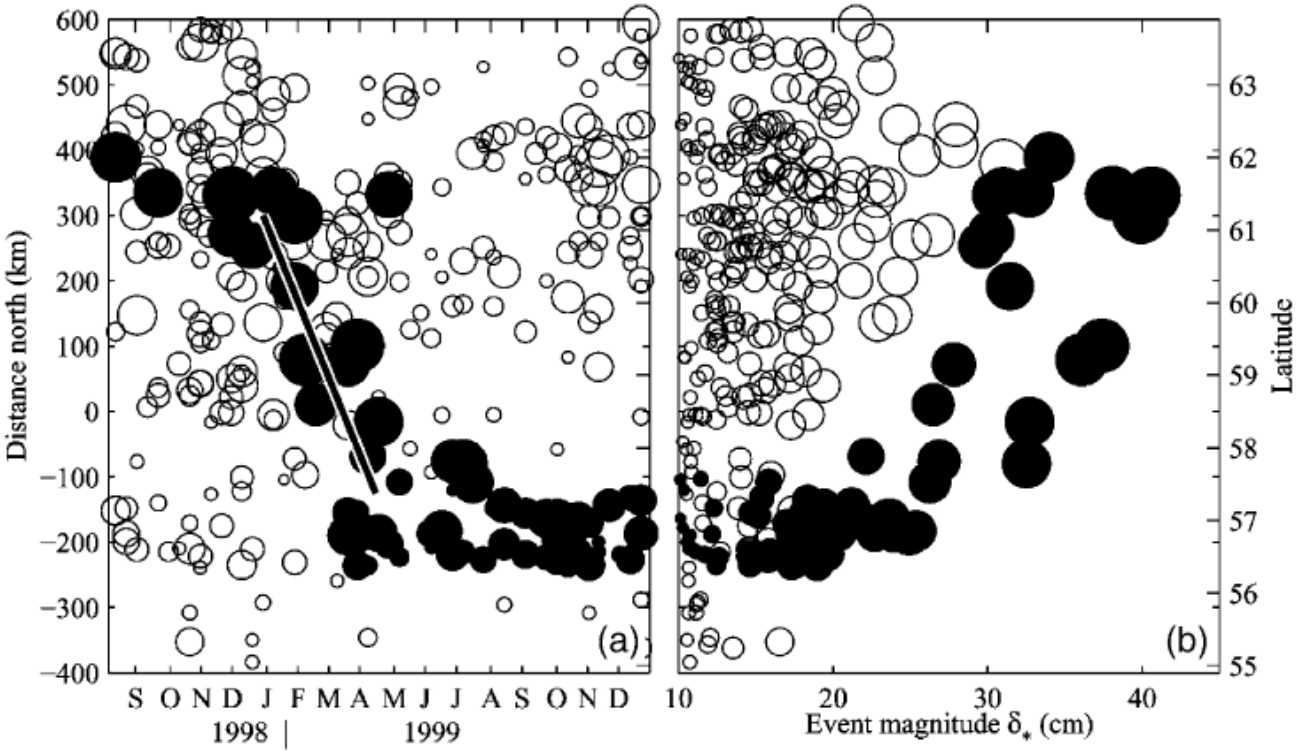
The Composite Event Plot



Then you can talk about a typical event by averaging to form a composite.

Here vertical bars are used to represent the spread of the ensemble about the composite.

3D Scatter Plot



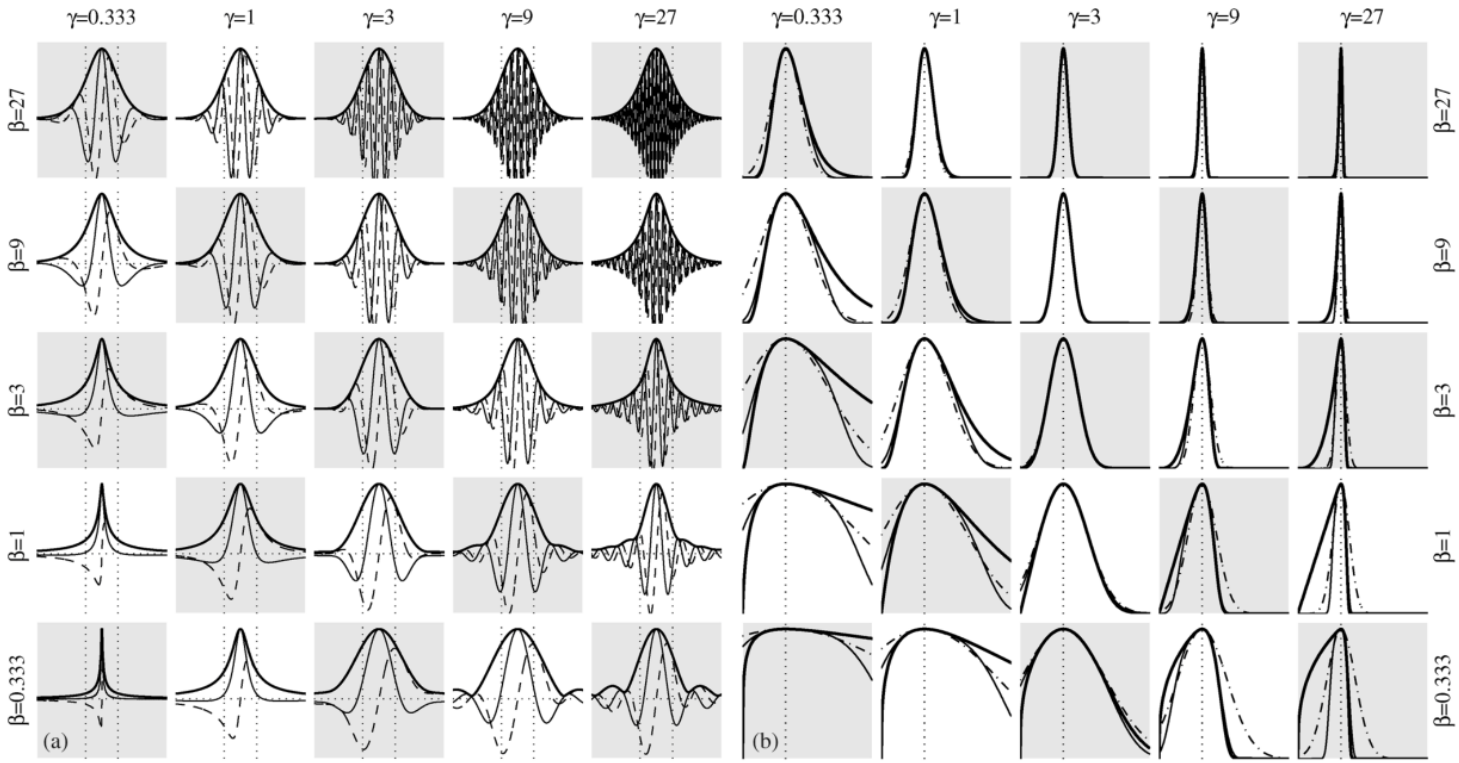
Quantitative 3D scatter plots can be very informative.

Size represents event magnitude, with the key shown on the right.
Note the use of (i) dovetailing axes and (ii) redundant information.

An example from

Lilly, J. M. and S. C. Olhede (2012). Generalized Morse wavelets as a superfamily of analytic wavelets. {link}

The Zoo Plot



In the Zoo Plot, we try to show a range of possibilities through numerous examples, as a function of some parameter space.

Here, left is time domain and right is Fourier domain.

Principles of Data Analysis, and Anti-Principles

Idea of principles and anti-principles taken from

Changing the Conversation: The 17 Principles of Conflict Resolution by Dana Caspersen

Principle: Get to know the data on its own terms, as you would get to know a person

Principle: Get to know the data on its own terms, as you would get to know a person

Anti-principle: Approach the data with expectations, hoping it will be what you want it to be

Principle: Keep your eyes sharp. Be on the lookout for unexpected patterns, mysteries, and problems

Principle: Keep your eyes sharp. Be on the lookout for unexpected patterns, mysteries, and problems

Anti-principle: Single-mindedly focus on one aspect of the data and ignore all others

Principle: Examine the data with the simplest possible methods at first, and gradually add complexity as needed

Principle: Examine the data with the simplest possible methods at first, and gradually add complexity as needed

Anti-principle: Begin by applying a fancy method you don't really understand using code that somebody else wrote

Principle: Keep your mind open to alternate interpretations and other possible explanations

Principle: Keep your mind open to alternate interpretations and other possible explanations

Anti-principle: Unconsciously assume that the first idea that came to you must be the right one

Principle: Try to hold multiple points of view simultaneously

Principle: Try to hold multiple points of view simultaneously

Anti-principle: Be content with the default view, the one you currently hold

Principle: Find a way to look at the data in such a way that it “collapses” or simplifies, revealing structure

Principle: Find a way to look at the data in such a way that it "collapses" or simplifies, revealing structure

Anti-principle: Look at the data in such a way that it becomes less compact or more complicated, obscuring structure

Principle: Follow your hunches, intuition, and curiosity tenaciously until you reach a definitive conclusion

Principle: Follow your hunches, intuition, and curiosity tenaciously until you reach a definitive conclusion

Anti-principle: Abandon hunches at the first sign of discouragement or difficulty. Be paralyzed by “black hat” thinking

Principle: As evidence accumulates for a particular hypothesis, try to critique it more strongly, looking for holes

Principle: As evidence accumulates for a particular hypothesis, try to critique it more strongly, looking for holes

Anti-principle: Let your favorite hypothesis get an “easy pass”. View the evidence in best possible light

Principle: As the story solidifies, honestly assess the limits of what can be proven, what is suggested, and what is not clear

Principle: As the story solidifies, honestly assess the limits of what can be proven, what is suggested, and what is not clear

Anti-principle: Overreach your conclusions, and feel the wrath of the reviewers. Understate, and lose impact

Principle: Learn to see the natural silhouette of a unit of scientific progress

Principle: Learn to see the natural silhouette of a unit of scientific progress

Anti-principle: Try to say too much, or too little, without telling a story

Principle: Come to terms with the fact that all units of scientific progress leave unanswered questions

Principle: Come to terms with the fact that all units of scientific progress leave unanswered questions

Anti-principle: Strive to answer all possible questions, and therefore never finish anything

Principle: Tell the easy or “low-hanging fruit” story first, and then build on it

Principle: Tell the easy or “low-hanging fruit” story first, and then build on it

Anti-principle: Get lost in possibilities, or obsessed with the difficult stories

Skillful qualities: Objective, impartial, open, cautious, creative, nimble, fluid, determined, prudent, meticulous

Data Analysis Roadmap

Phase 0: Preparation

Being like an accountant.

- **Studying coding basics and best practices**
- **Gathering unprocessed datasets**
  - Calibration & quality control
  - Despiking as needed (visual/subjective; global or running $n\sigma$ deviation, median, 2nd difference, etc.; iterative exclusion; fancier methods)
  - Interpolation to uniform grid if needed (linear, quadratic, or spline; optimal interpolation, local polynomial fit, or spline basis; error assessment)
  - Dataset organization & documentation
  - ⇒ Working dataset version
    (! With all processing code in a script !)
- **Gathering other relevant processed data**
- **Meta-organization**

Phase 1: Exploratory

Being like an explorer getting a first glimpse of new lands.

- Looking at the data
- Making a bunch of plots (! Don't be afraid to make hardcopies !)
- Following intuition & having fun
- Clearing your mind of preconceptions
- Not paying attention to the literature at all
- Sticking with very simple methods (simple statistics, 1D and 2D histograms and statistics, line plots, simple smoothing)
- Sidestepping minor technical problems; flag these instead
- Keeping eyes out for suspicious or intriguing features
- Refactoring: abstracting figure types, code blocks, etc.
- ⇒ Data report document / figure stack
  (! With all processing code in a script / Jupyter notebook!)
- ⇒ Qualitative, intuitive assessment of noise vs. signal
- ⇒ List of features or aspect worthy of further investigation
- ⇒ ? Possibly iterate to Phase 0 with new information

Do not work directly on the command line without saving your code! You will be stuck in data purgatory for all eternity!

Phase 2: Investigating

Being like a detective patiently building a case.

- Brainstorming (with pen and paper) interpretations of the data
- Forming **multiple** hypotheses to explain interesting features
- Countering physical hypotheses with a suitable null hypothesis (e.g. noise or artifacts)
- Sticking with simple methods (see next slide)
- Sidestepping roadblocks (! Do not stop when you hit obstacles!)
- Gathering evidence in support of / opposing these hypotheses
- Setting aside personal preferences (yours and your advisor's)
- Setting aside what everyone believes to be true
- Maintaining a curious, open mind
- Building a case through plots, argumentation, and analysis
- Keeping in mind the limitations of the dataset
- Asking: Is the evidence conclusive?
- Asking: What other datasets / perspectives could be helpful?
- Asking: What other methods may be called for?

Phase 2 Methods

- Clarifying variability at relevant timescales: diurnal, tidal, inertial, annual, etc.
- Separating variability with simple smoothing, harmonic fits, formation of composite cycle (e.g. annual), etc., *and residual*
- Examining all of the Phase 1 aspects on separated components
- Studying theory of relevant processes to familiarize yourself
- Forming simple conceptual, kinematic, or statistical model for the observed features
- Gathering ancillary or environmental data for potential forcing, causative, or associated processes
- Correlations, EOFs, etc.
- Higher-order or circular statistics

Phase 3: Forensics

Bringing in specialize methods to help the investigation.

- Be aware that this is often not necessary!
- Talk to colleagues with more experience
- Ask: do I want to learn this, or instead, find a collaborator?
- Be prepared to sit down and study for weeks or months
- Learn the method thoroughly *before* applying it to your dataset!

Some possibilities:

- Fourier spectral analysis
- Wavelet analysis
- Stochastic modeling
- Interpolation methods: OI, local polynomial fitting, spline
- Correlation analysis: CCA, MCA, MLR, SVD
- Clustering methods
- Statistical hypotheses testing

- ⇒ Definitive evidence *or* inconclusive results
- ⇒ Proceed to Phase 4 *or* iterate with Phase 2

Phase 4: Closing the Case

Being like a lawyer presenting the case to the jury.

- Understanding your results within the context of the literature
- Assessing what makes a unit of scientific progress
- Putting together a case with figures, equations, and arguments
- Considering all possible objections
  (! Use Thinking Hats and role-playing!)
- Iterating to earlier phases as needed
- Building a watertight case
- Being honest about the limitations of your results
- Seeing unanswered questions as future possibilities
- Knowing when to stop
- Remembering to not get personally involved with your client

Also!

- Learning about data formats and conventions; iterate to 0 and 1

⇒ A scientific paper; a finished, shared dataset; open software