The use of imprecise processing to improve accuracy in weather & climate prediction

Peter Düben, Tim Palmer, Hugh McNamara

University of Oxford

What is imprecise processing?

For this talk a bit-reproducible double precision floating point operation is called a **precise operation**.

What is imprecise processing?

For this talk a bit-reproducible double precision floating point operation is called a **precise operation**.

Basic idea:
If we sacrifice the dogma that every operation needs to be calculated precisely we can reduce computational costs.

What is imprecise processing?

For this talk a bit-reproducible double precision floating point operation is called a **precise operation**.

Basic idea:
If we sacrifice the dogma that every operation needs to be calculated precisely we can reduce computational costs.

Two approaches to imprecise processing:

1. **Stochastic processors**
   Reduction of precision due to faulty calculations, but significant reduction of power consumption.
2. **Low floating point precision (double $\rightarrow$ single $\rightarrow$ ...)**
   Increased performance of model simulations since less storage is needed and more data fits into memory and cache.

Outline

1. Bit representation and stochastic processors
2. Imprecise processing in weather and climate models
3. Numerical tests with a dynamical core: IGCM
4. Conclusion and Outlook

A short introduction to bit representation

- The computer represents the integer number 102090 as a string of 32 bits (00000000000000011001110110011010).

A short introduction to bit representation

- The computer represents the integer number 102090 as a string of 32 bits (00000000000000011001110110010 10).
- Each bit represents a power of two:

$$102090 = 0\cdot 2^0 + 1\cdot 2^1 + 0\cdot 2^2 + 1\cdot 2^3 + 0\cdot 2^4 + 0\cdot 2^5 + 1\cdot 2^6 \ldots = \sum_{i=0}^{31} b_i 2^i$$

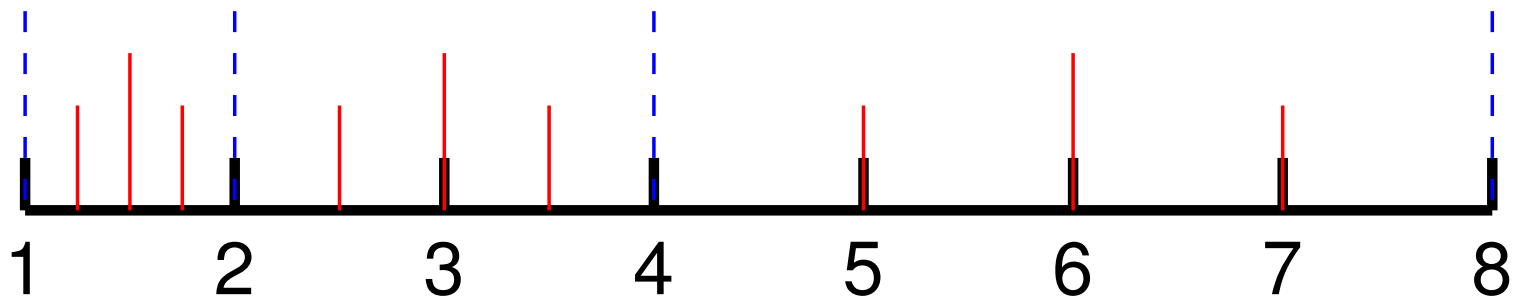
A short introduction to bit representation

- The computer represents the integer number 102090 as a string of 32 bits (00000000000000011001110110010 10).
- Each bit represents a power of two:

$$102090 = 0\cdot 2^0 + 1\cdot 2^1 + 0\cdot 2^2 + 1\cdot 2^3 + 0\cdot 2^4 + 0\cdot 2^5 + 1\cdot 2^6 \ldots = \sum_{i=0}^{31} b_i 2^i$$

- A real number a is represented as a 64 bit floating point number:

$$a = (-1)^S \left(1 + \sum_{i=1}^{52} b_{-i} 2^{-i}\right) 2^E, \quad \text{where} \quad E = \left(\sum_{i=0}^{10} e_i 2^i\right) - 1023.$$



Tuning a processor

- **A reduction of the supplied voltage** can reduce the power consumption of the processor. $\rightarrow$ Voltage scaling.
- If we reduce the supplied voltage, more time is needed.
- **A reduction of the wall clock time** can increase the performance of the processor.

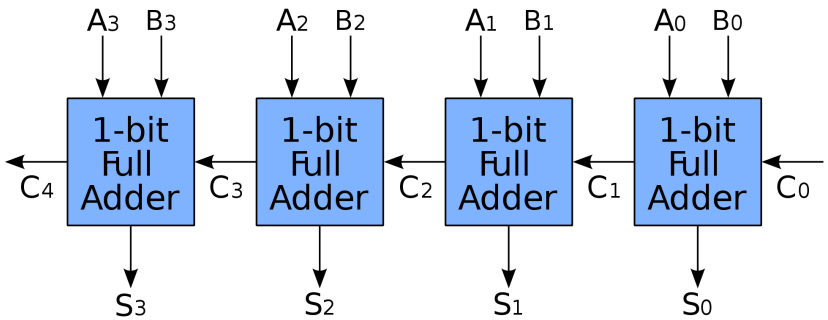


Figure from wikipedia.com

Tuning a processor

- **A reduction of the supplied voltage** can reduce the power consumption of the processor. $\rightarrow$ Voltage scaling.
- If we reduce the supplied voltage, more time is needed.
- **A reduction of the wall clock time** can increase the performance of the processor.

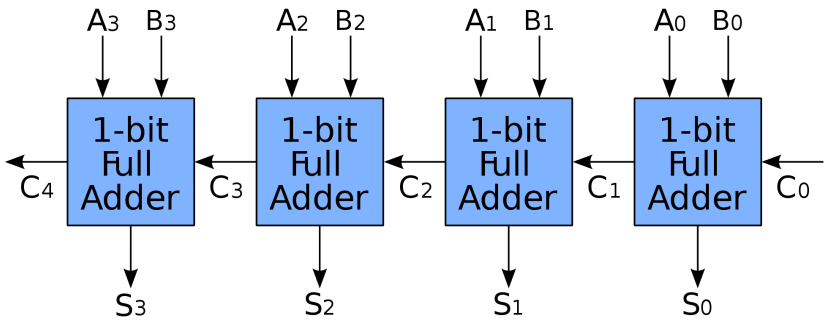


Figure from wikipedia.com

Today's computers are designed to be exact in all operations.

Tuning a processor

- **A reduction of the supplied voltage** can reduce the power consumption of the processor. $\rightarrow$ Voltage scaling.
- If we reduce the supplied voltage, more time is needed.
- **A reduction of the wall clock time** can increase the performance of the processor.

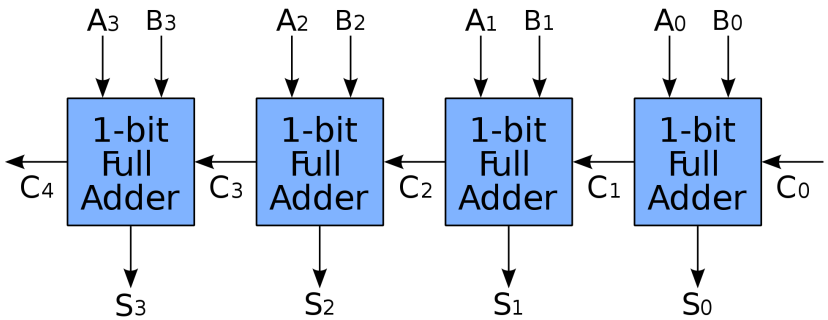


Figure from wikipedia.com

Today's computers are designed to be exact in all operations.

If we can cope with some errors we can be faster and/or cheaper.

Tuning a processor

$$\begin{array}{r} 0001 \\ +0001 \\ \hline 0010 \end{array} \qquad \begin{array}{r} 0101 \\ +0011 \\ \hline 1000 \end{array}$$

Tuning a processor

- Different operations need a different time to be calculated correctly.

  $\rightarrow$ Operations have to wait for the wall clock time: Timing slack.

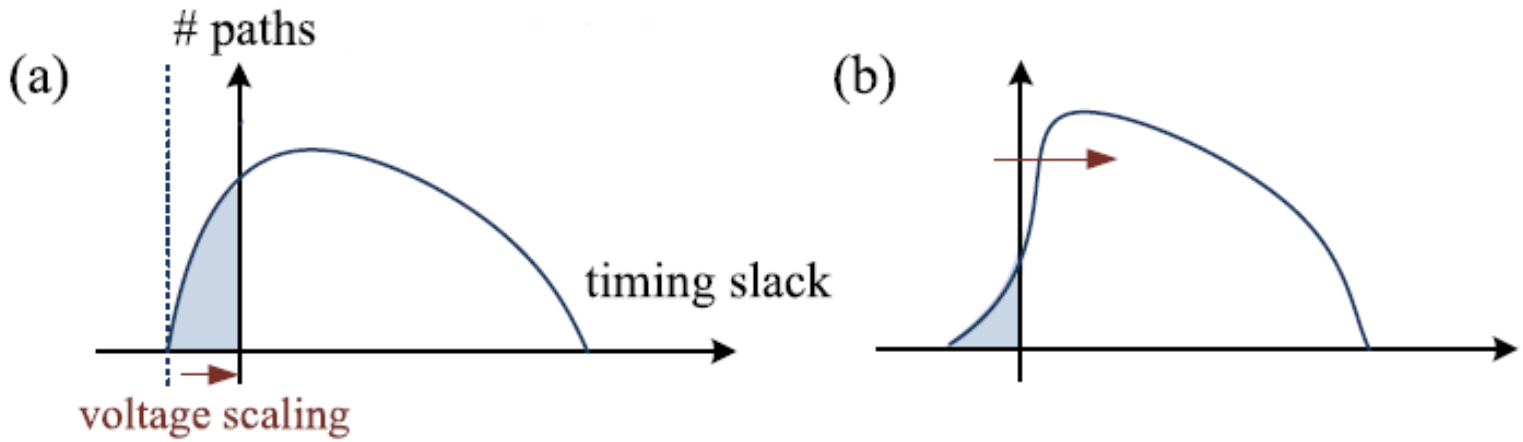
- The required time for each operation is fixed by the architecture of the processor.

Tuning a processor

- Different operations need a different time to be calculated correctly.

  $\rightarrow$ Operations have to wait for the wall clock time: Timing slack.

- The required time for each operation is fixed by the architecture of the processor.



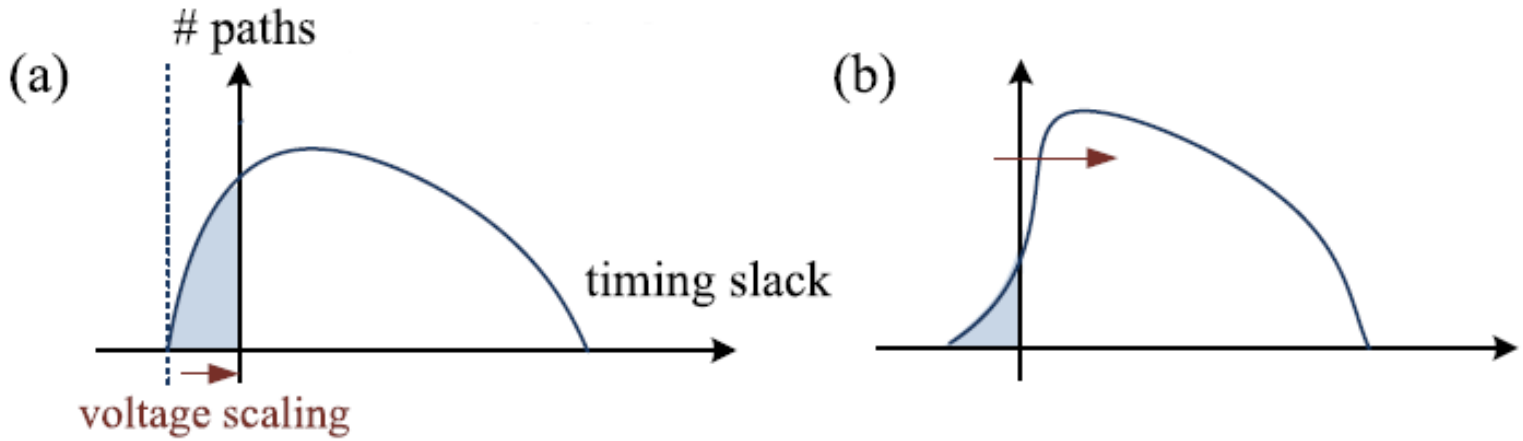
Sartori et al. (2011)

Tuning a processor

- Different operations need a different time to be calculated correctly.

  $\rightarrow$ Operations have to wait for the wall clock time: Timing slack.

- The required time for each operation is fixed by the architecture of the processor.



Sartori et al. (2011)

The error rate can be reduced massively, if the architecture is changed $\rightarrow$ Stochastic Processor.

Stochastic processors: Fast and/or cheap with errors

But is it really worth it?

Stochastic processors: Fast and/or cheap with errors

But is it really worth it?

Recent studies discuss a decrease of power consumption by...

- ...12-20% at a 1-2% error rate
  (Kahng et al. 2010, Sartori et al. 2011)
- ...up to 90% at a 10% error rate
  (Lingamneni et al. 2012)

Stochastic processors: Fast and/or cheap with errors

But is it really worth it?

Recent studies discuss a decrease of power consumption by...

- ...12-20% at a 1-2% error rate
  (Kahng et al. 2010, Sartori et al. 2011)
- ...up to 90% at a 10% error rate
  (Lingamneni et al. 2012)

Energy demand and error resilience are two of the main challenges towards an exascale super computer.

Stochastic processors: Fast and/or cheap with errors

But is it really worth it?

Recent studies discuss a decrease of power consumption by...

- ...12-20% at a 1-2% error rate
  (Kahng et al. 2010, Sartori et al. 2011)
- ...up to 90% at a 10% error rate
  (Lingamneni et al. 2012)

Energy demand and error resilience are two of the main challenges towards an exascale super computer.

Stochastic processors could reduce the computational cost and enable to increase the resolution.?

Imprecise processing in weather and climate models

- Spectral models allow to treat short and large scales independently.
- The large scales need to be calculated exactly.
- The small scales can not be solved exactly anyway (viscosity, parametrisation,...).
- The small scales are expensive:
  - T21: $N = 242$
  - T31: $N = 512$
  - T42: $N = 924$.

  N is the number of coefficients needed to represent a horizontal scalar field.

Imprecise processing in weather and climate models

- Spectral models allow to treat short and large scales independently.
- The large scales need to be calculated exactly.
- The small scales can not be solved exactly anyway (viscosity, parametrisation,...).
- The small scales are expensive:
  - T21: $N = 242$
  - T31: $N = 512$
  - T42: $N = 924$.

  N is the number of coefficients needed to represent a horizontal scalar field.

We should try to use stochastic processors to compute the small scales (Palmer 2012).

Imprecise processing in weather and climate models

- Spectral models allow to treat short and large scales independently.
- The large scales need to be calculated exactly.
- The small scales can not be solved exactly anyway (viscosity, parametrisation,...).
- The small scales are expensive:
  - T21: $N = 242$
  - T31: $N = 512$
  - T42: $N = 924$.

  N is the number of coefficients needed to represent a horizontal scalar field.

We should try to use stochastic processors to compute the small scales (Palmer 2012).

Higher resolution $\rightarrow$ less parametrisation.

Imprecise processing in weather and climate models

- Spectral models allow to treat short and large scales independently.
- The large scales need to be calculated exactly.
- The small scales can not be solved exactly anyway (viscosity, parametrisation,...).
- The small scales are expensive:
  - T21: $N = 242$
  - T31: $N = 512$
  - T42: $N = 924$.

  N is the number of coefficients needed to represent a horizontal scalar field.

We should try to use stochastic processors to compute the small scales (Palmer 2012).

Higher resolution $\rightarrow$ less parametrisation.

It might be possible to perform hardware ensembles.

What are the limits?

- The computer represents the integer number 102090 as a string of 32 bits (00000000000000011001110110010 10).
- Each bit represents a power of two:

$$102090 = 0\cdot 2^0 + 1\cdot 2^1 + 0\cdot 2^2 + 1\cdot 2^3 + 0\cdot 2^4 + 0\cdot 2^5 + 1\cdot 2^6 \ldots = \sum_{i=0}^{31} b_i 2^i$$

- A real number a is represented as a 64 bit floating point number:

$$a = (-1)^S \left(1 + \sum_{i=1}^{52} b_{-i} 2^{-i}\right) 2^E, \quad \text{where} \quad E = \left(\sum_{i=0}^{10} e_i 2^i\right) - 1023.$$

What are the limits?

- The computer represents the integer number 102090 as a string of 32 bits (00000000000000011001110110010 10).
- Each bit represents a power of two:

$$102090 = 0\cdot 2^0 + 1\cdot 2^1 + 0\cdot 2^2 + 1\cdot 2^3 + 0\cdot 2^4 + 0\cdot 2^5 + 1\cdot 2^6 \ldots = \sum_{i=0}^{31} b_i 2^i$$

- A real number a is represented as a 64 bit floating point number:

$$a = (-1)^S \left(1 + \sum_{i=1}^{52} b_{-i} 2^{-i}\right) 2^E, \quad \text{where} \quad E = \left(\sum_{i=0}^{10} e_i 2^i\right) - 1023.$$

We only touch the 52 bits of the significand of floating point numbers!!!

An emulated stochastic processor

The bad news:

- Stochastic processors are not produced yet.
- We do not know the precise properties of a stochastic processor.

The good news:

- We can emulate the use of stochastic processors by flipping bits with a prescribed error rate.
- Our tests provide an application for the stochastic processor community and can influence their development towards the demands of the climate community.

How to emulate stochastic processors

- We emulate an error in x percent of the floating point operations, where x is the error rate.
- Floating point operations are summations, multiplications, divisions, sine, cosine,... .
- We emulate an error by switching one bit of the significand. We set the probability of a switch to be the same for all bits.

A dynamical core: IGCM

IGCM (or the Reading spectral model) is a simplified global circulation model that...

- ...is based on spectral discretisation methods.
- ...solves the primitive equations on the sphere.
- ...uses σ coordinates.

IGCM on a stochastic processor

What happens if we calculate the whole model on an emulated stochastic processor with 10% error rate?

IGCM on a stochastic processor

What happens if we calculate the whole model on an emulated stochastic processor with 10% error rate?

It crashes!

IGCM on a stochastic processor

What happens if we calculate the whole model on an emulated stochastic processor with 10% error rate?

It crashes!

Lets have a look at a possible scale separation.

IGCM: More details

One time step in IGCM consists of...

1. **... a Legendre Transformation:**
   Spherical harmonics $\rightarrow$ Fourier series at each latitude
2. **... a Fast Fourier Transformation:**
   Fourier series at each latitude $\rightarrow$ Grid point space
3. **... the calculation of non-linear tendencies**
4. **... a Fast Fourier Transformation:**
   Grid point space $\rightarrow$ Fourier series at each latitude
5. **... a Legendre Transformation:**
   Fourier series at each latitude $\rightarrow$ Spherical harmonics
6. **... the time stepping scheme**

IGCM: A closer look

Parts of the model in which scale separation is possible:

- The Legendre transformation.
- The time step.

Parts of the model in which scale separation is not possible:

- The Fast Fourier Transformation.
- The calculation of the non-linear tendencies.

IGCM: A closer look

Parts of the model in which scale separation is possible:

- The Legendre transformation (LT).
- The time step (Time step).

Parts of the model in which scale separation is not possible:

- The Fast Fourier Transformation (FFT).
- The calculation of the non-linear tendencies (Non-linear).

Numerical work load:

Resolution	LT & Time step	FFT	Non-linear
T21	41%	35%	23%
T31	45%	35%	20%
T42	48%	33%	19%
T84	64%	25%	11%

IGCM: A closer look

Parts of the model in which scale separation is possible:

- The Legendre transformation (LT).
- The time step (Time step).

Parts of the model in which scale separation is not possible:

- The Fast Fourier Transformation (FFT).
- The calculation of the non-linear tendencies (Non-linear).

Numerical work load:

Resolution	LT & Time step	FFT	Non-linear
T21	41%	35%	23%
T31	45%	35%	20%
T42	48%	33%	19%
T84	64%	25%	11%

Scale separation can not be performed in large parts of the model.

The FFT on a stochastic processor

What happens if the whole FFT is performed on a stochastic processor?


Exact


0.54 % error


2.05 % error


7.58 % error

Lingamneni et al. (2012)

IGCM: Simulations with inexact arithmetic

We perform simulations in which we use the stochastic processor to calculate...

Case 1: ...the non-linear tendencies.

$\rightarrow$ 18 % of the floating point operations

IGCM: Simulations with inexact arithmetic

We perform simulations in which we use the stochastic processor to calculate...

Case 1: ...the non-linear tendencies.

$\rightarrow$ 18 % of the floating point operations

Case 2: ...the non-linear tendencies and the Legendre transformation and the time stepping scheme for the small wave lengths (T32-T42).

$\rightarrow$ 31 % of the floating point operations

IGCM: Simulations with inexact arithmetic

We perform simulations in which we use the stochastic processor to calculate...

Case 1: ...the non-linear tendencies.

$\rightarrow$ 18 % of the floating point operations

Case 2: ...the non-linear tendencies and the Legendre transformation and the time stepping scheme for the small wave lengths (T32-T42).

$\rightarrow$ 31 % of the floating point operations

Case 3: ...the non-linear tendencies, the Legendre transformation and the time stepping scheme for the small wave lengths (T32-T42) and the FFT.

$\rightarrow$ 84 % of the floating point operations

IGCM: Simulations with inexact arithmetic

We perform simulations in which we use the stochastic processor to calculate...

Case 1: ...the non-linear tendencies.

$\rightarrow$ 18 % of the floating point operations

Case 2: ...the non-linear tendencies and the Legendre transformation and the time stepping scheme for the small wave lengths (T32-T42).

$\rightarrow$ 31 % of the floating point operations

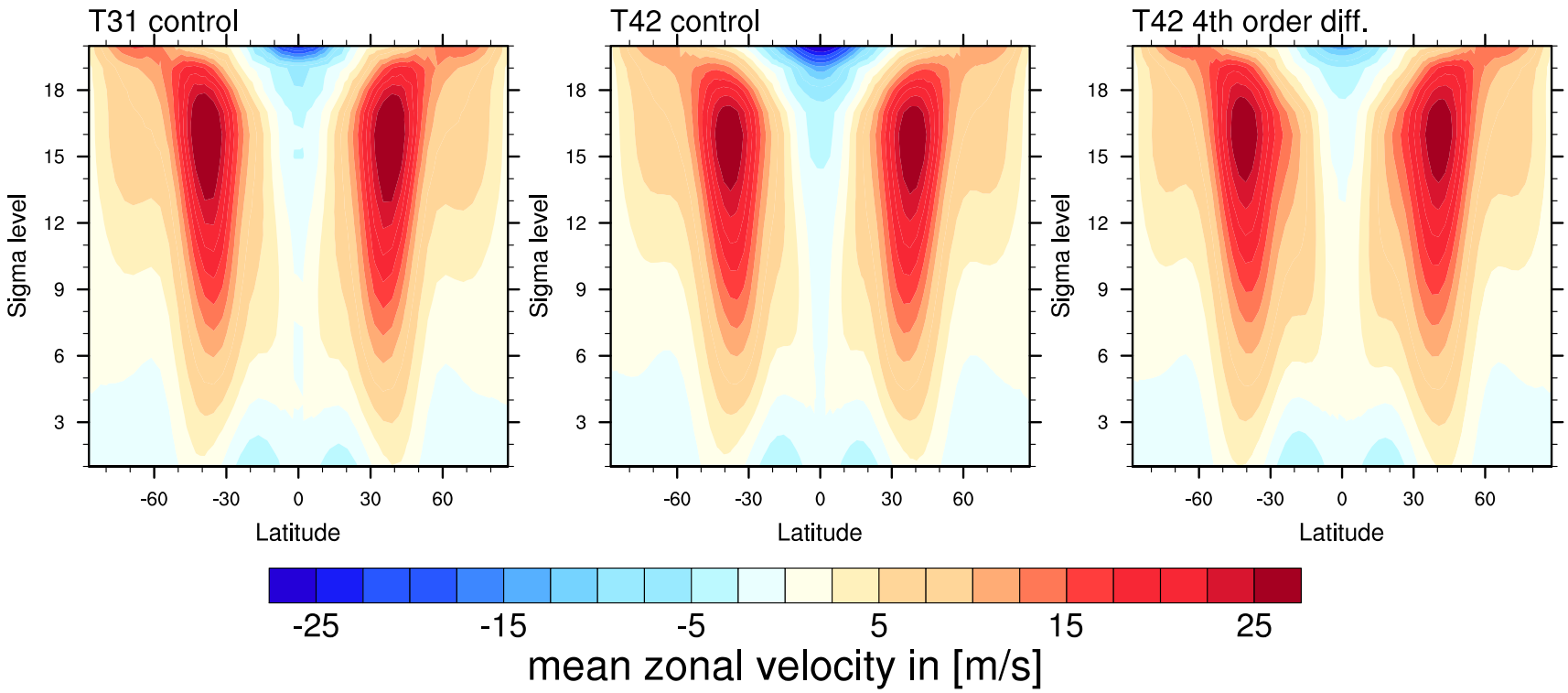
Case 3: ...the non-linear tendencies, the Legendre transformation and the time stepping scheme for the small wave lengths (T32-T42) and the FFT.

$\rightarrow$ 84 % of the floating point operations

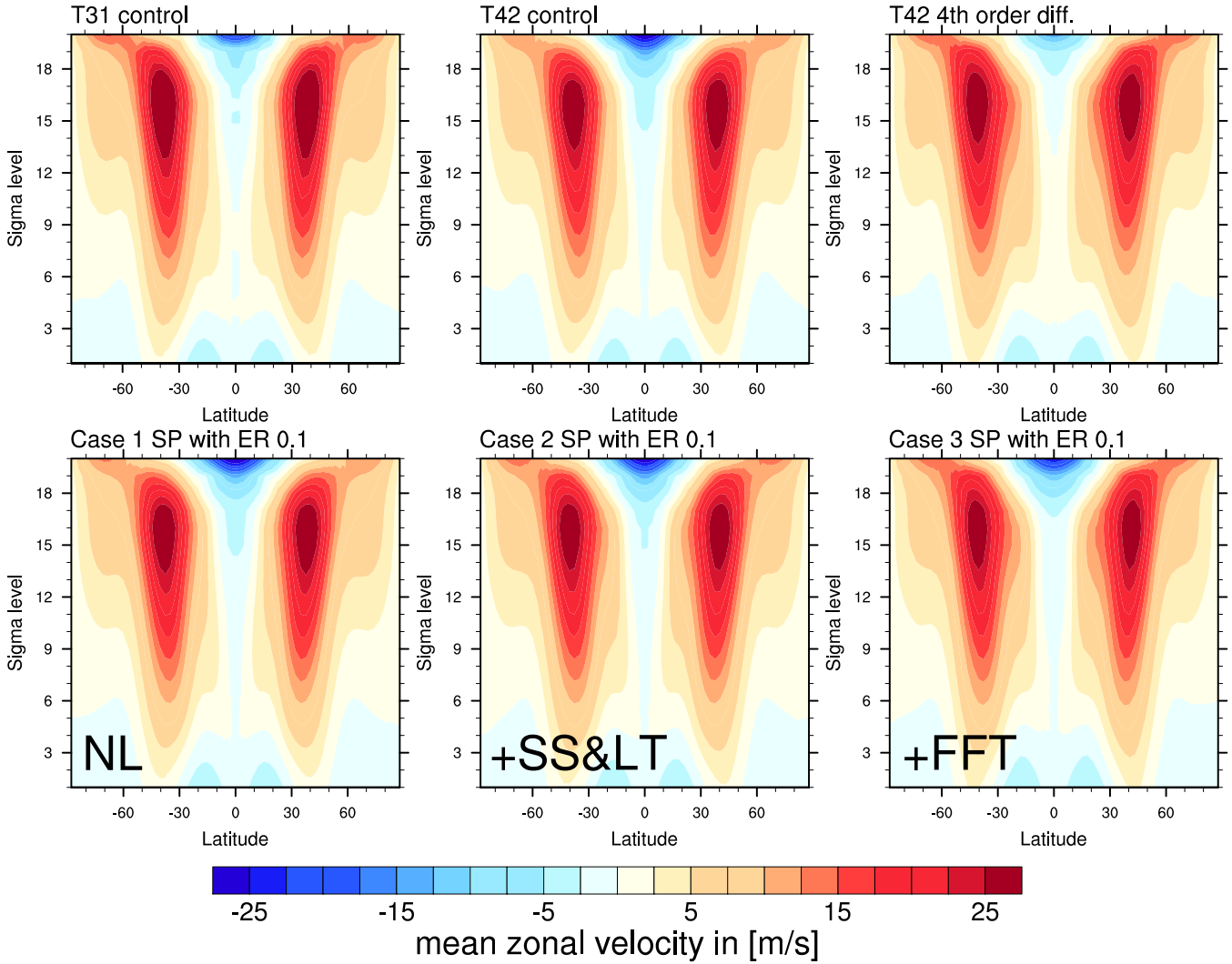
We emulate stochastic processors with an error rate of 1% or 10%.

The Held-Suarez test

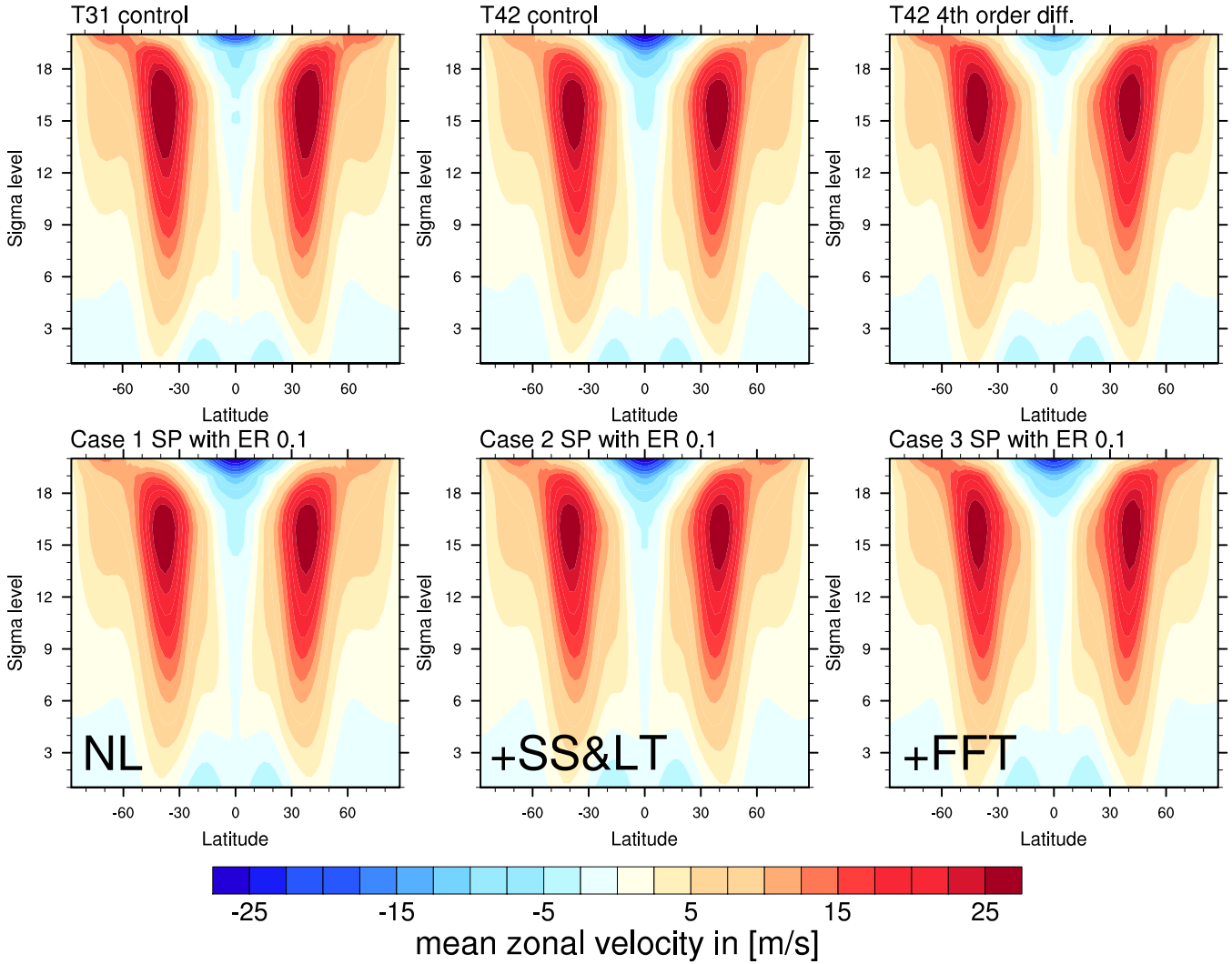
- The temperature field is relaxed to a zonally symmetric field.
- We simulate with 20 vertical levels and T31 or T42 resolution.
- We calculate the mean zonal velocity field from simulations over 10000 days.



IGCM: Scale separated simulations

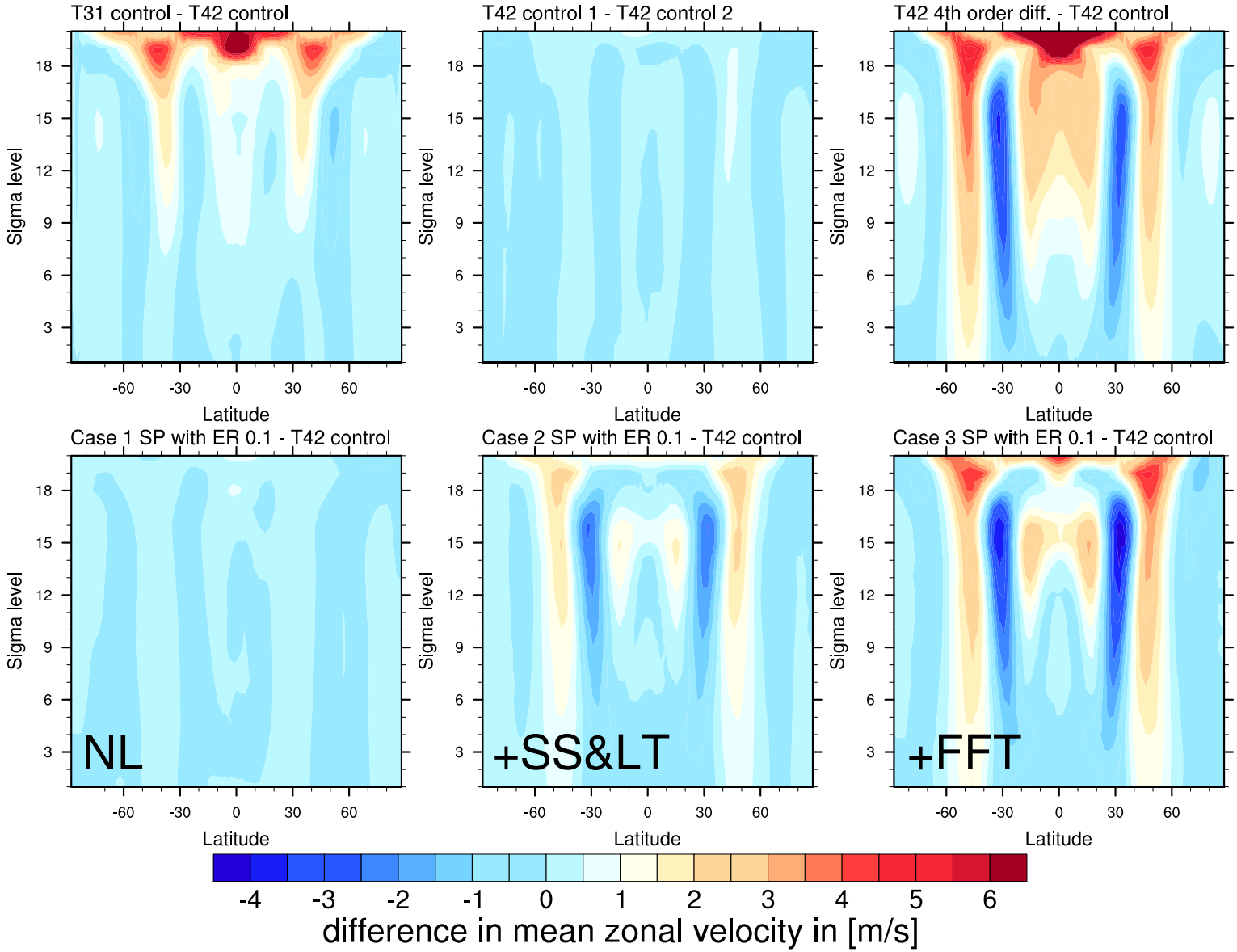


IGCM: Scale separated simulations



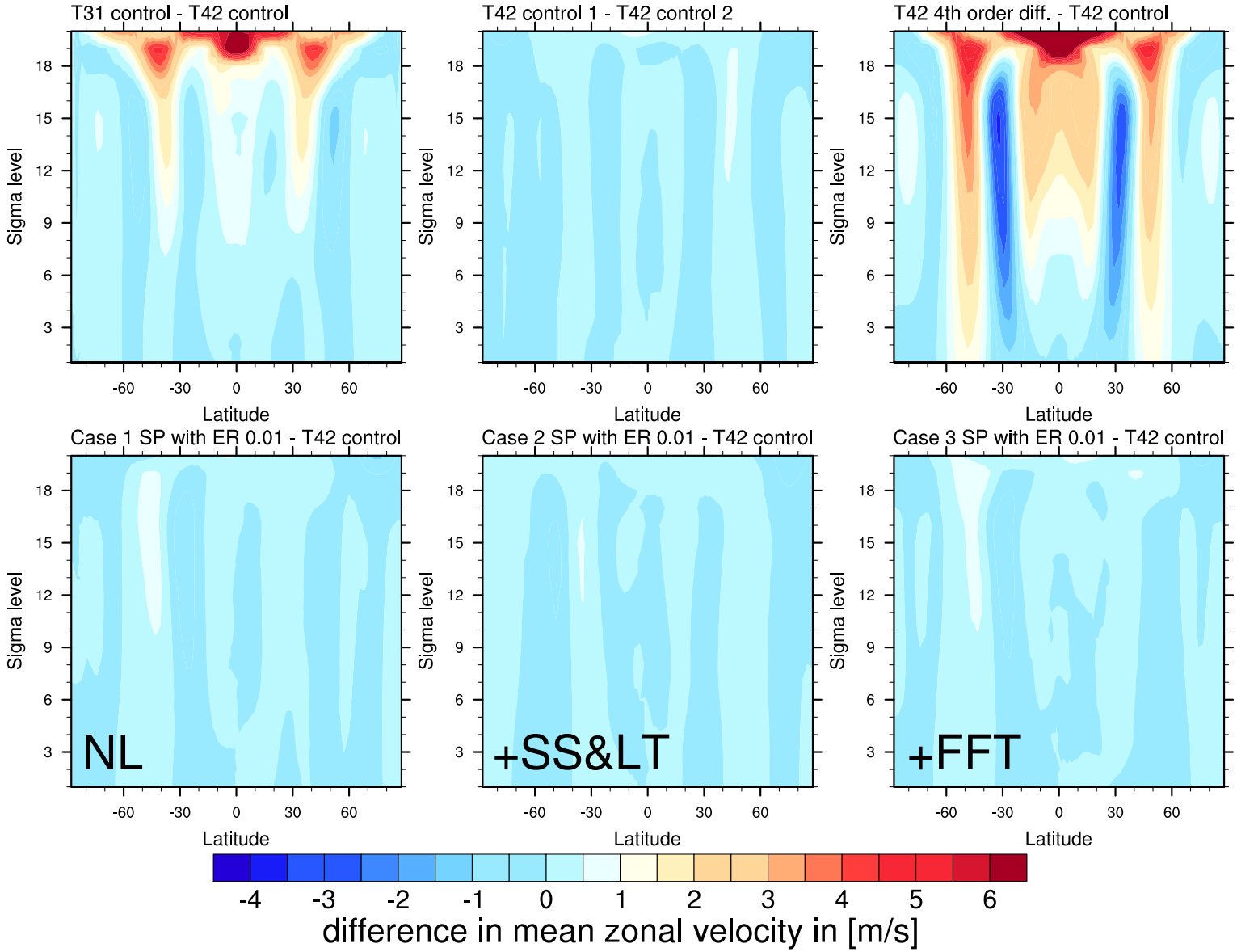
The results are promising.

IGCM: Scale separated simulations



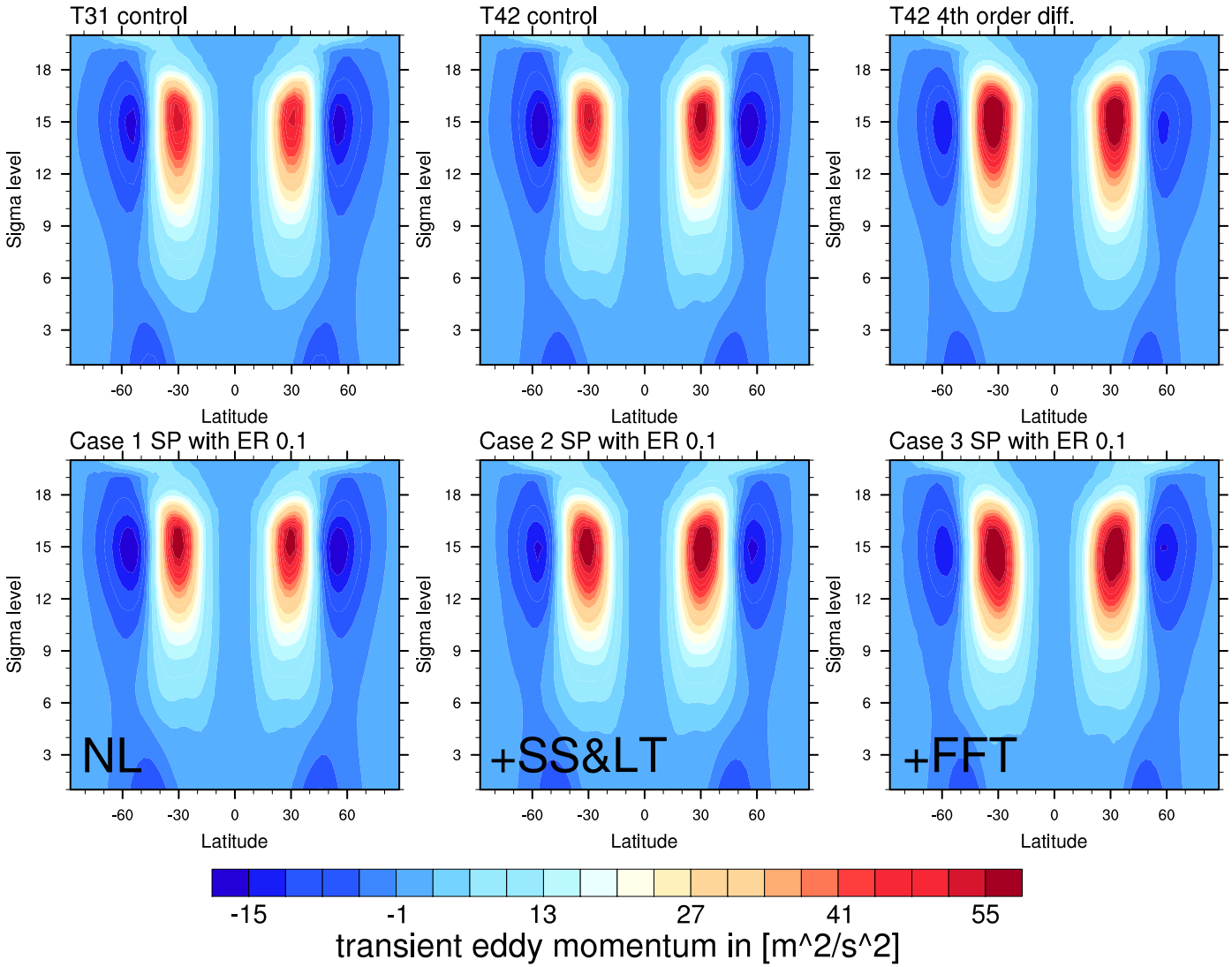
Difference plots with T42 control simulation. Error rate 10%.

IGCM: Scale separated simulations



Difference plots with T42 control simulation. Error rate 1%.

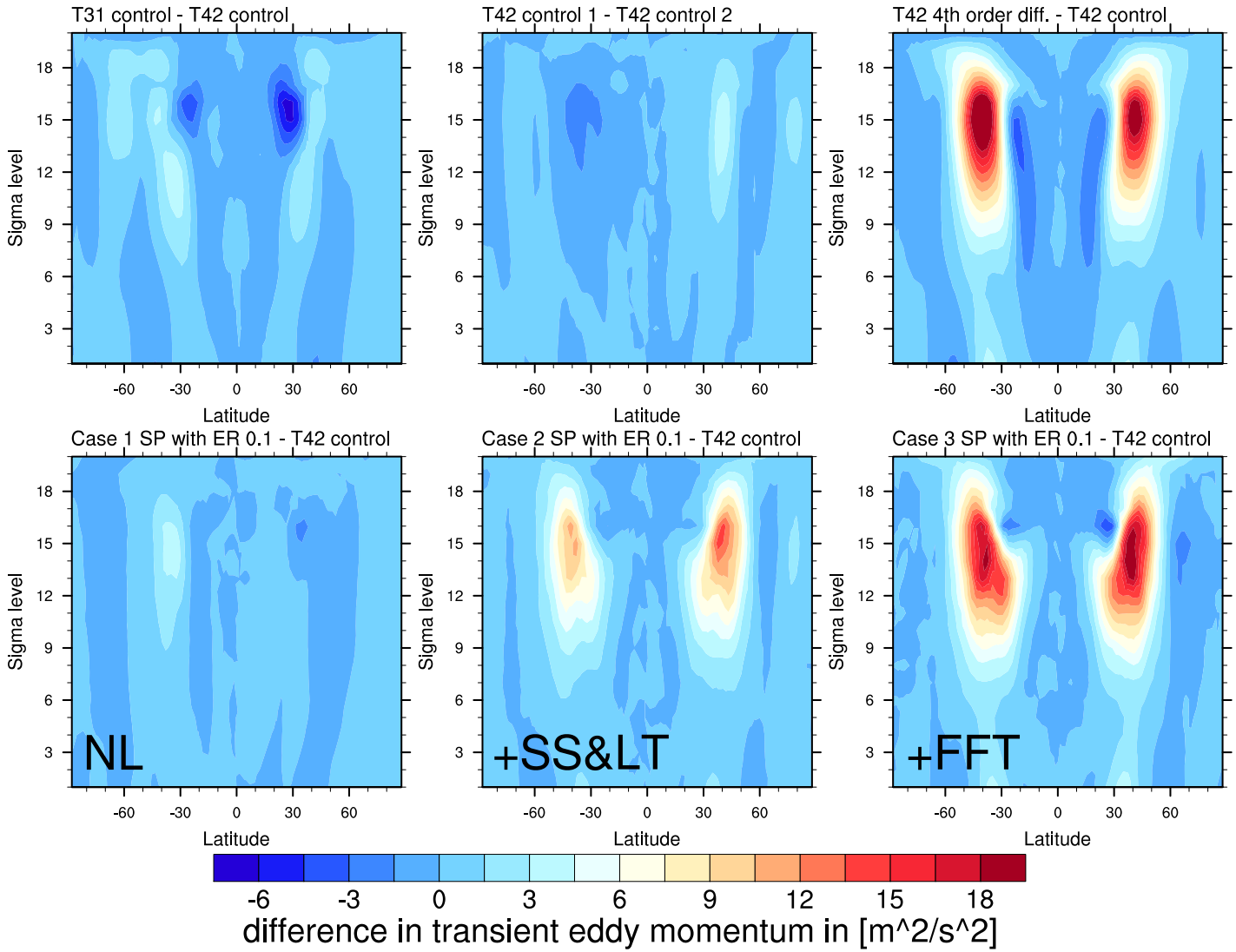
IGCM: Scale separated simulations



Transient eddy momentum $\overline{[u'^* v'^*]}$. Error rate 10%.

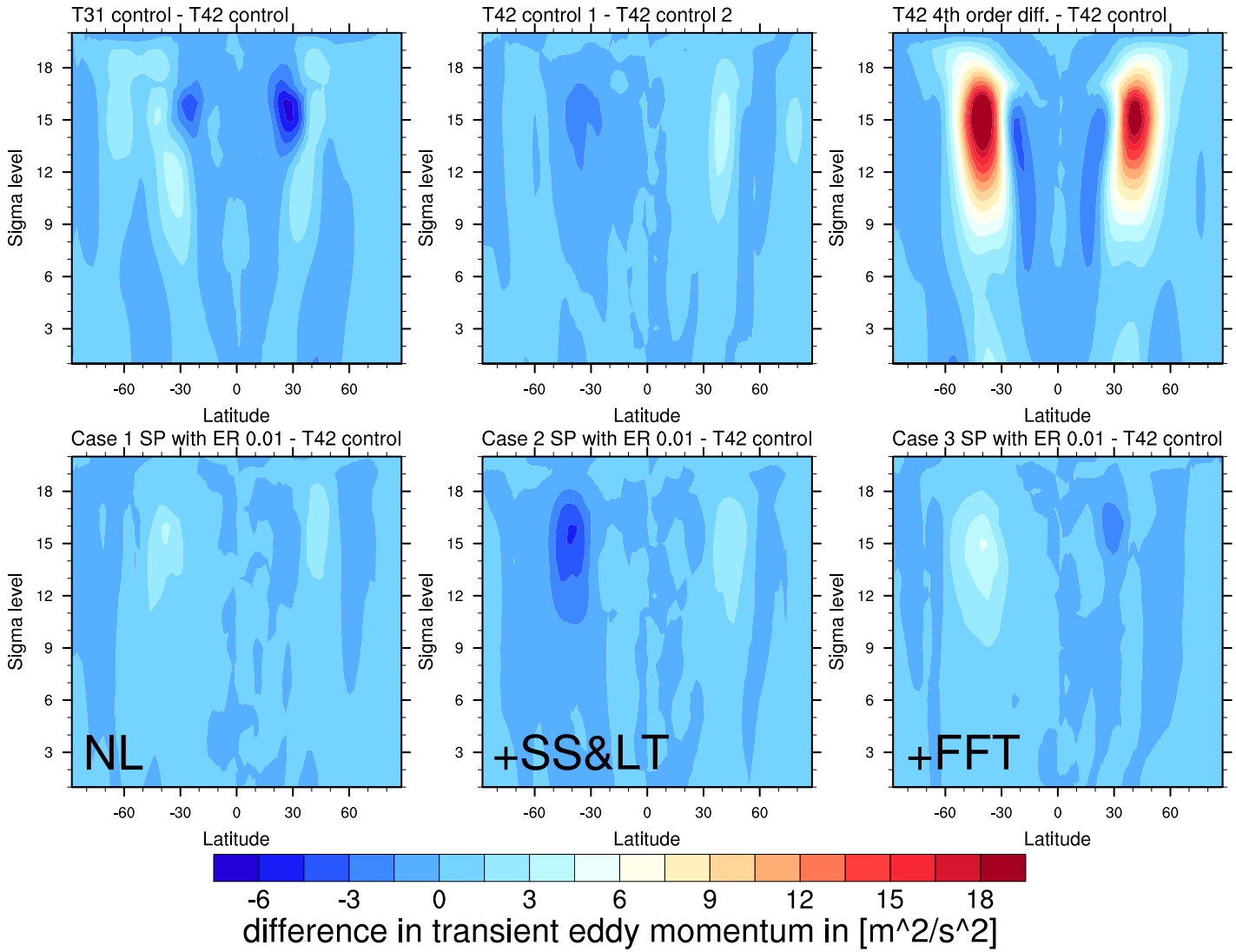
The prime denotes deviations from the time average, the asterix denotes deviations from the zonal average, the square brackets denote a zonal average, and the over-bar denotes a time average.

IGCM: Scale separated simulations



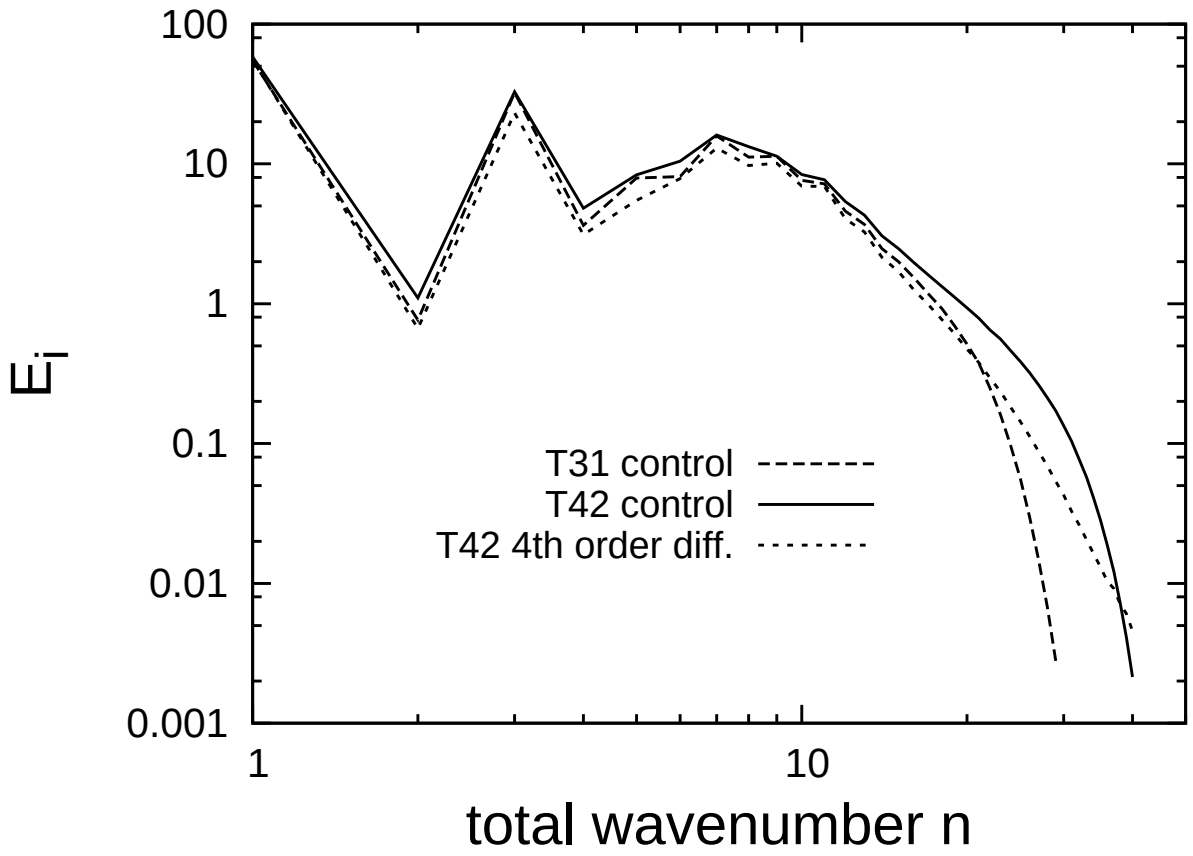
Difference plots with T42 control simulation. Error rate 10%.

IGCM: Scale separated simulations

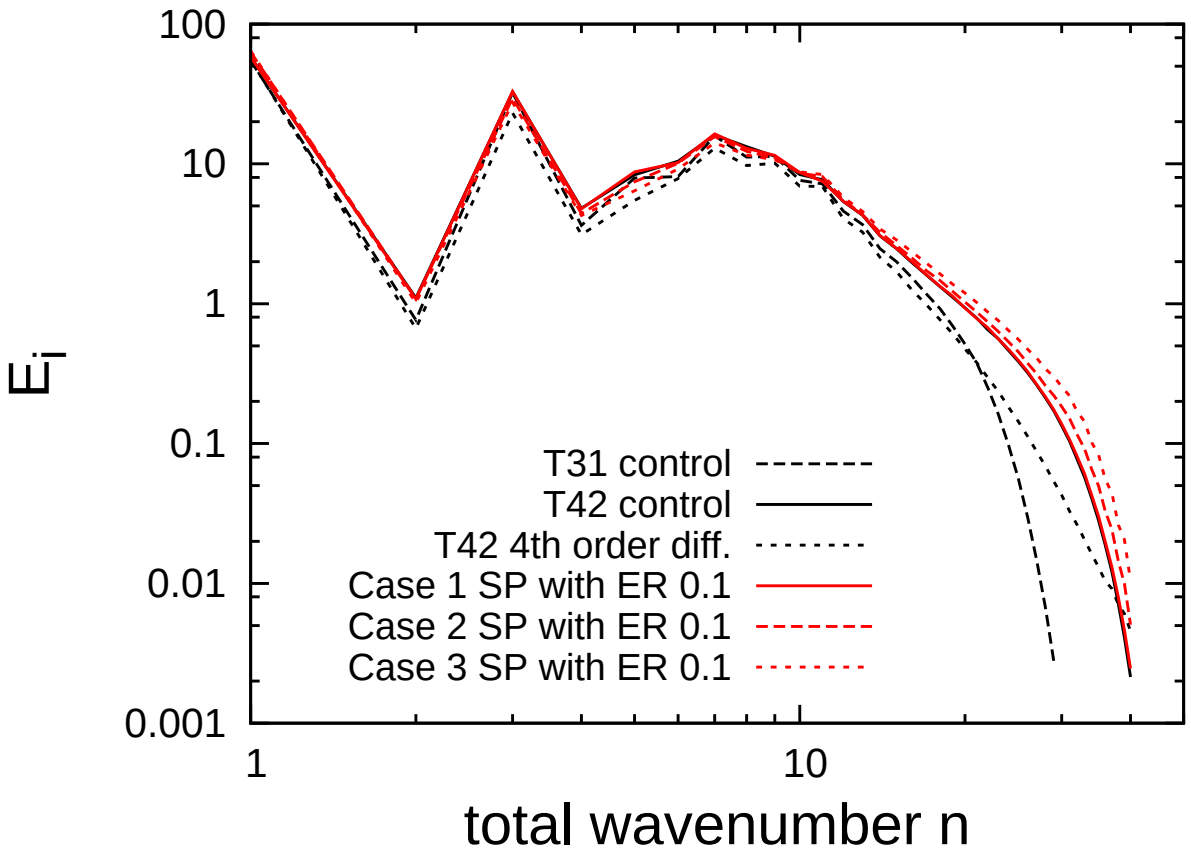


Difference plots with T42 control simulation. Error rate 1%.

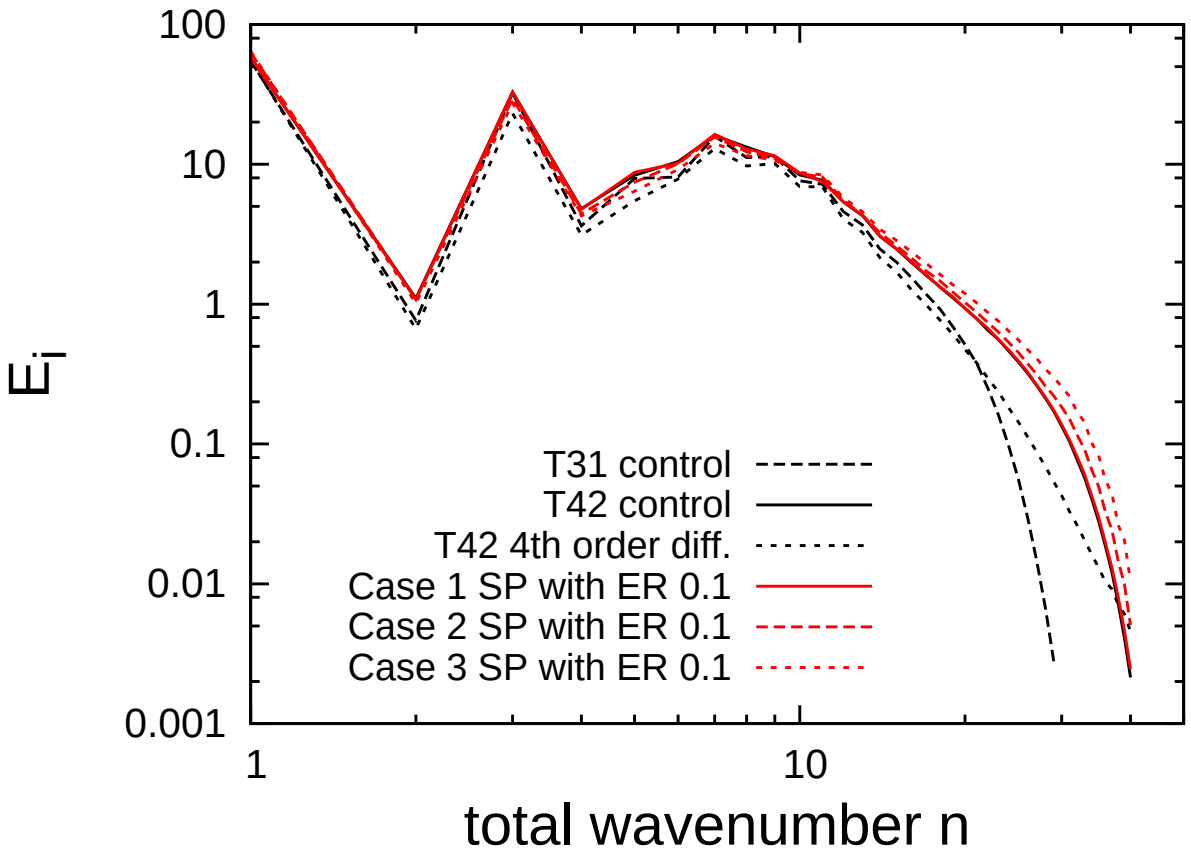
IGCM: Energy spectra



IGCM: Energy spectra



IGCM: Energy spectra



The energy spectra look OK with a slight introduction of energy to the small scales.

IGCM with lower floating point precision

- If we flip a specific bit in 50% of the operations we have a good 'emulator' of a calculation with decreased real number accuracy (double $\rightarrow$ single $\rightarrow$ ...).
- We performed the same test runs as before (case 1-3) with a **6 bits** representation for the significand of floating points.

IGCM with lower floating point precision

- If we flip a specific bit in 50% of the operations we have a good 'emulator' of a calculation with decreased real number accuracy (double $\rightarrow$ single $\rightarrow$ ...).
- We performed the same test runs as before (case 1-3) with a **6 bits** representation for the significand of floating points.

Scale separation is necessary!

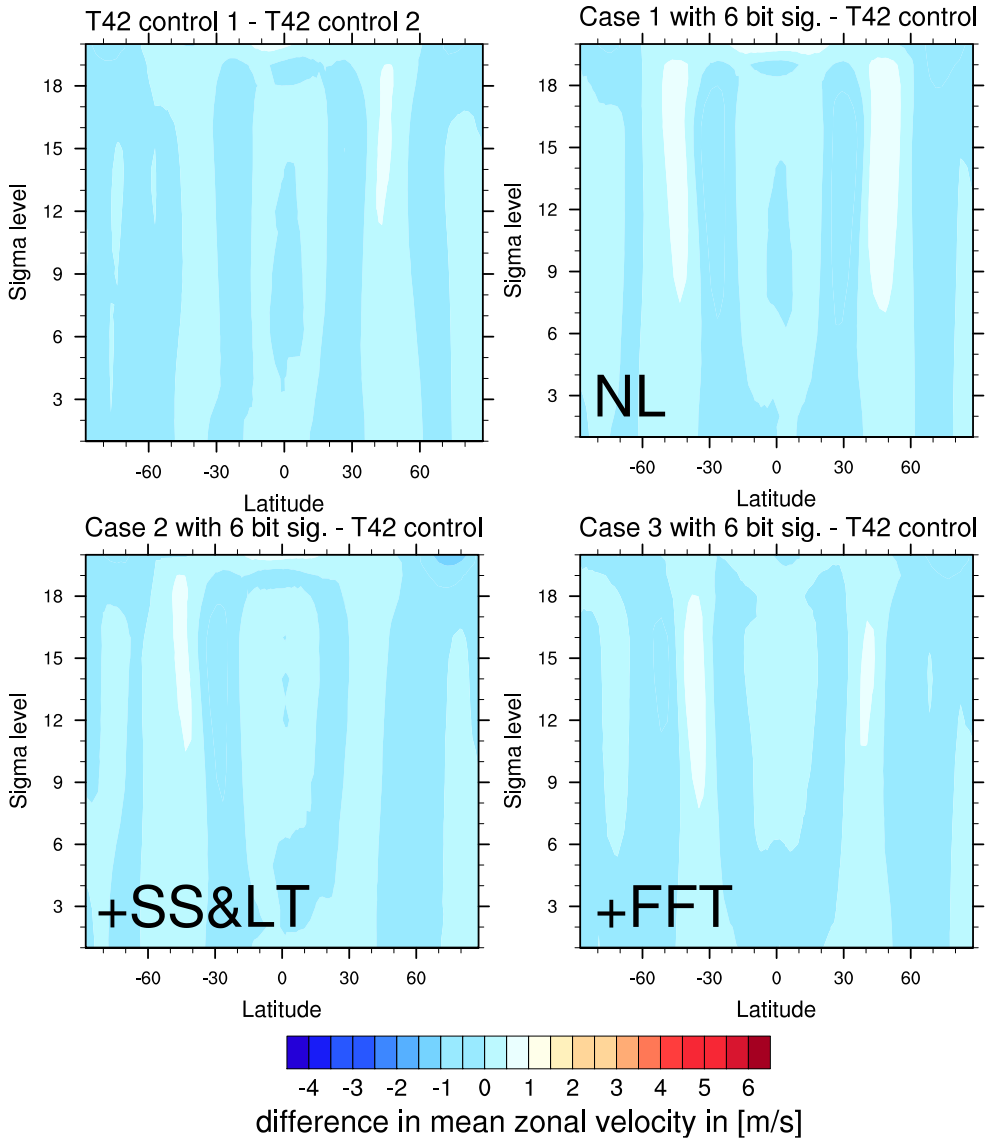
IGCM with lower floating point precision

- If we flip a specific bit in 50% of the operations we have a good 'emulator' of a calculation with decreased real number accuracy (double $\rightarrow$ single $\rightarrow$ ...).
- We performed the same test runs as before (case 1-3) with a **6 bits** representation for the significand of floating points.

Scale separation is necessary!

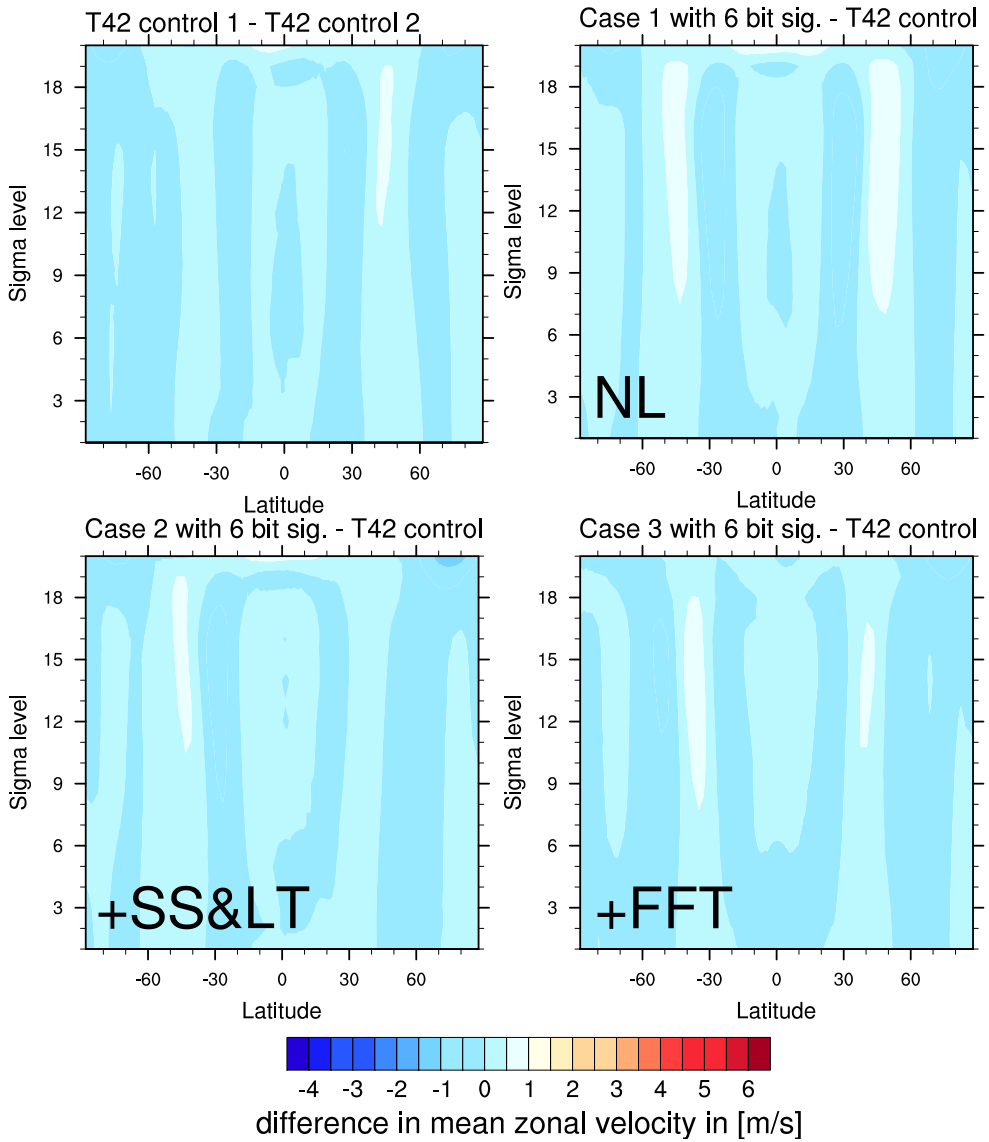
Simulations with a 4 bit significand crash. There are limits!

IGCM with lower floating point precision



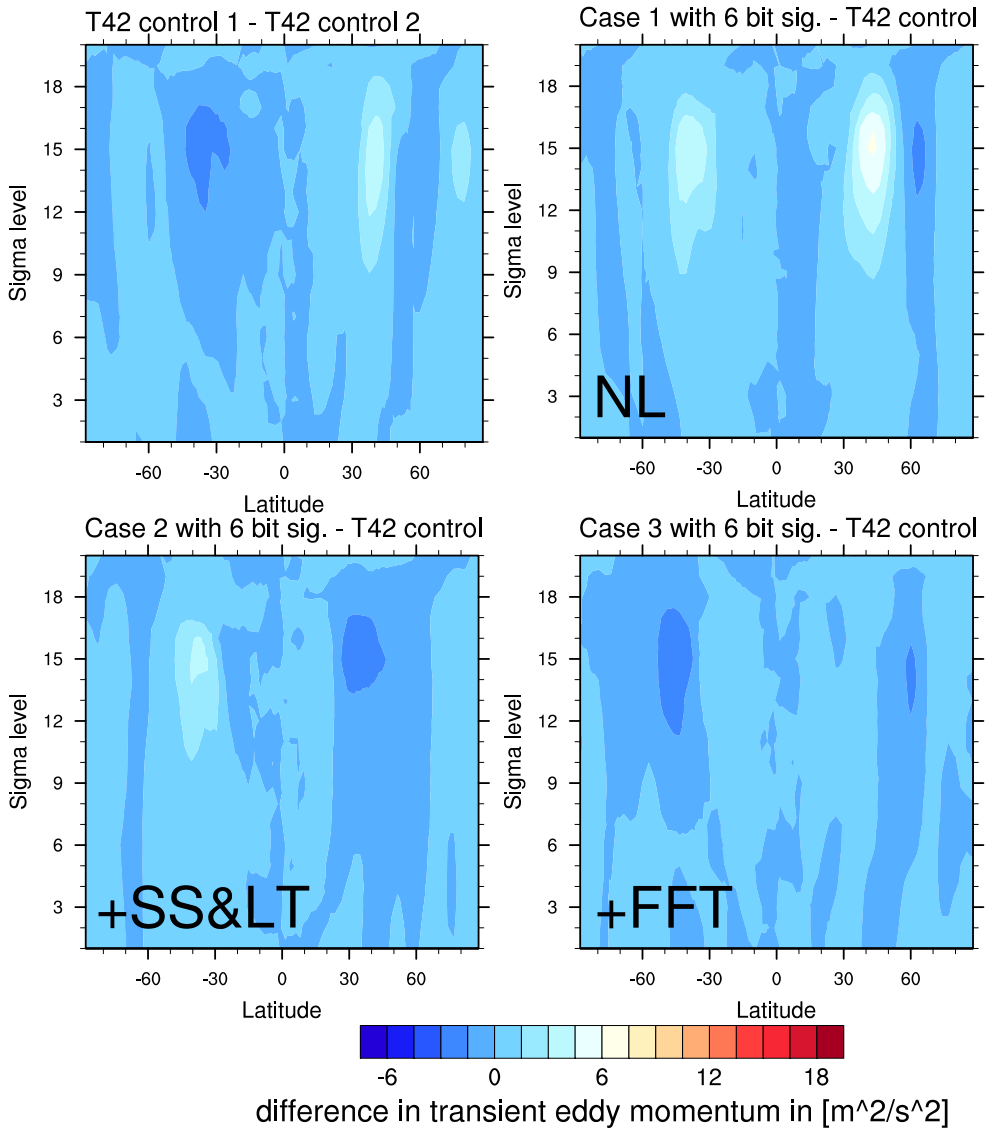
Difference plots for mean zonal velocity.

IGCM with lower floating point precision



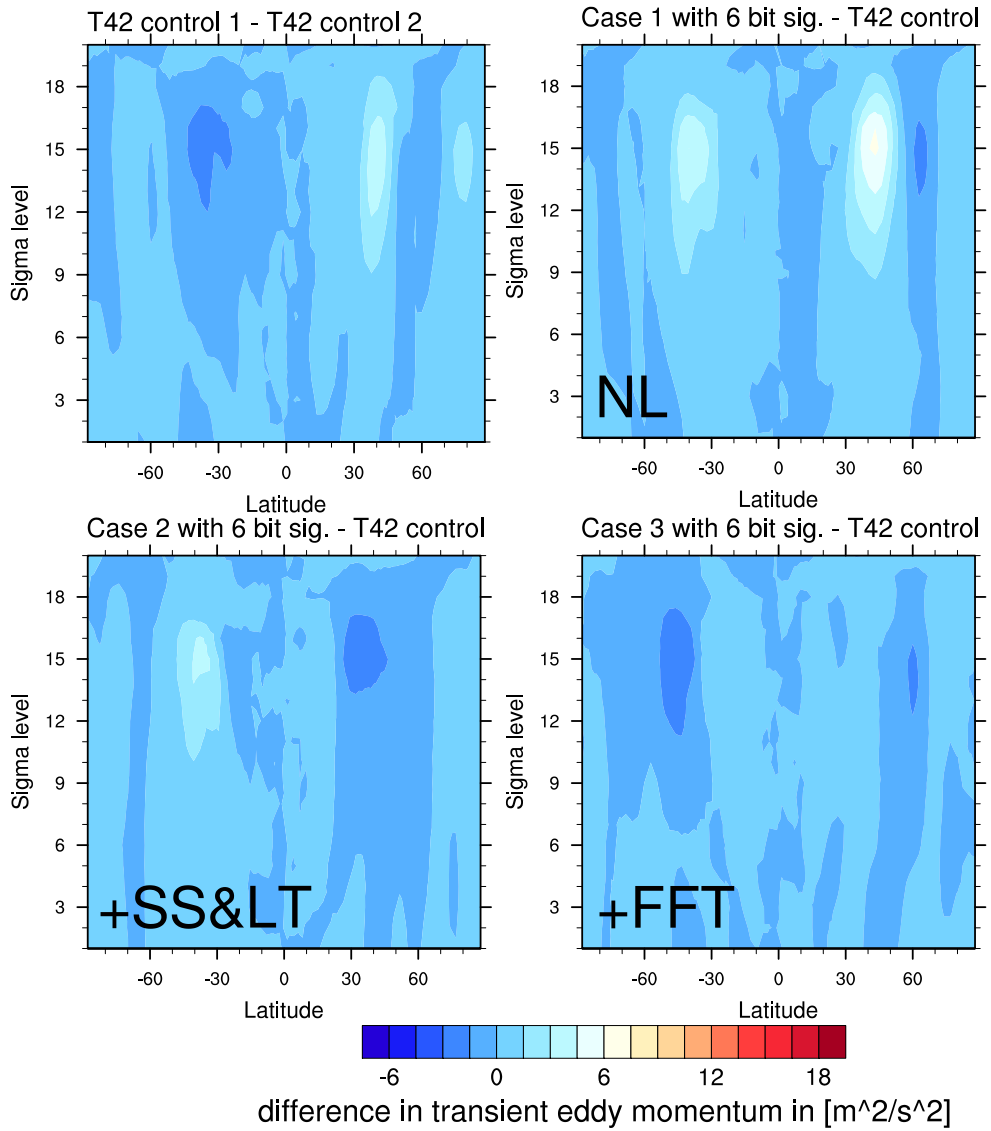
Promising!

IGCM with lower floating point precision



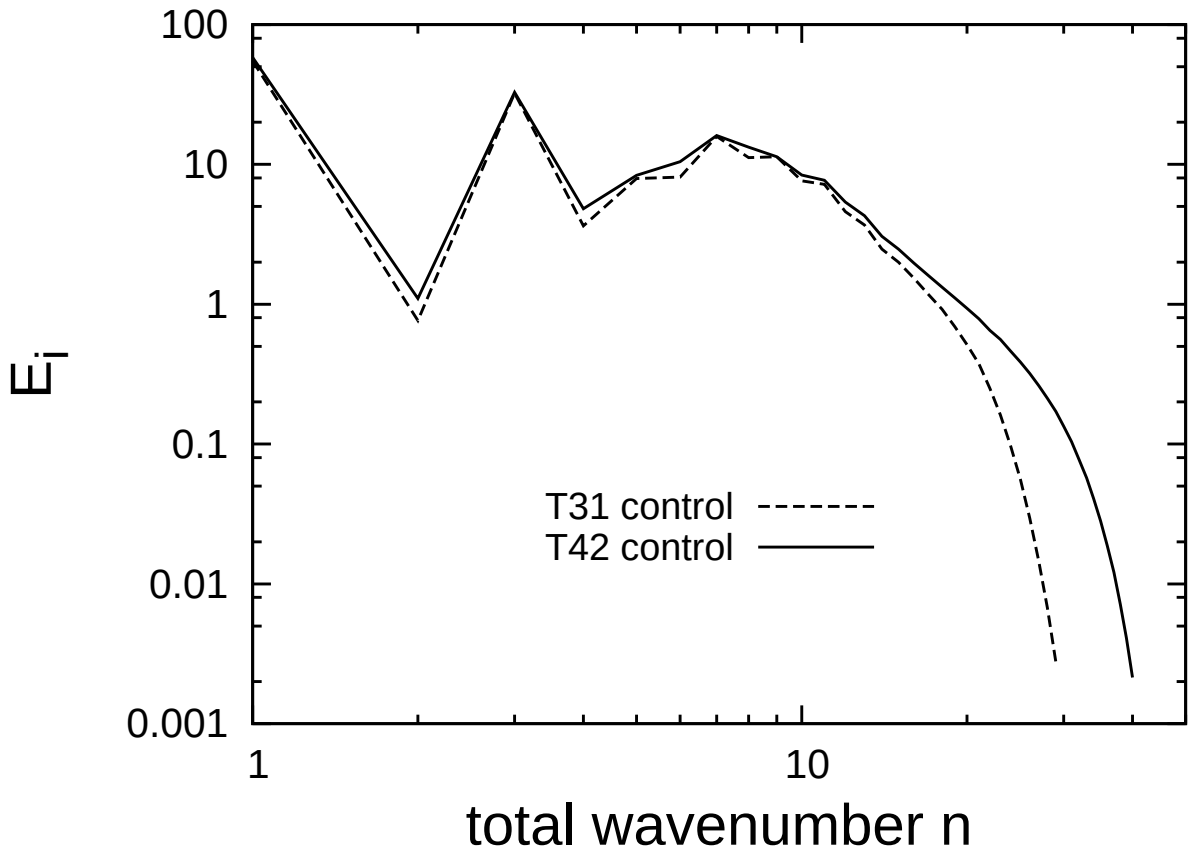
Difference plots for transient eddy momentum.

IGCM with lower floating point precision

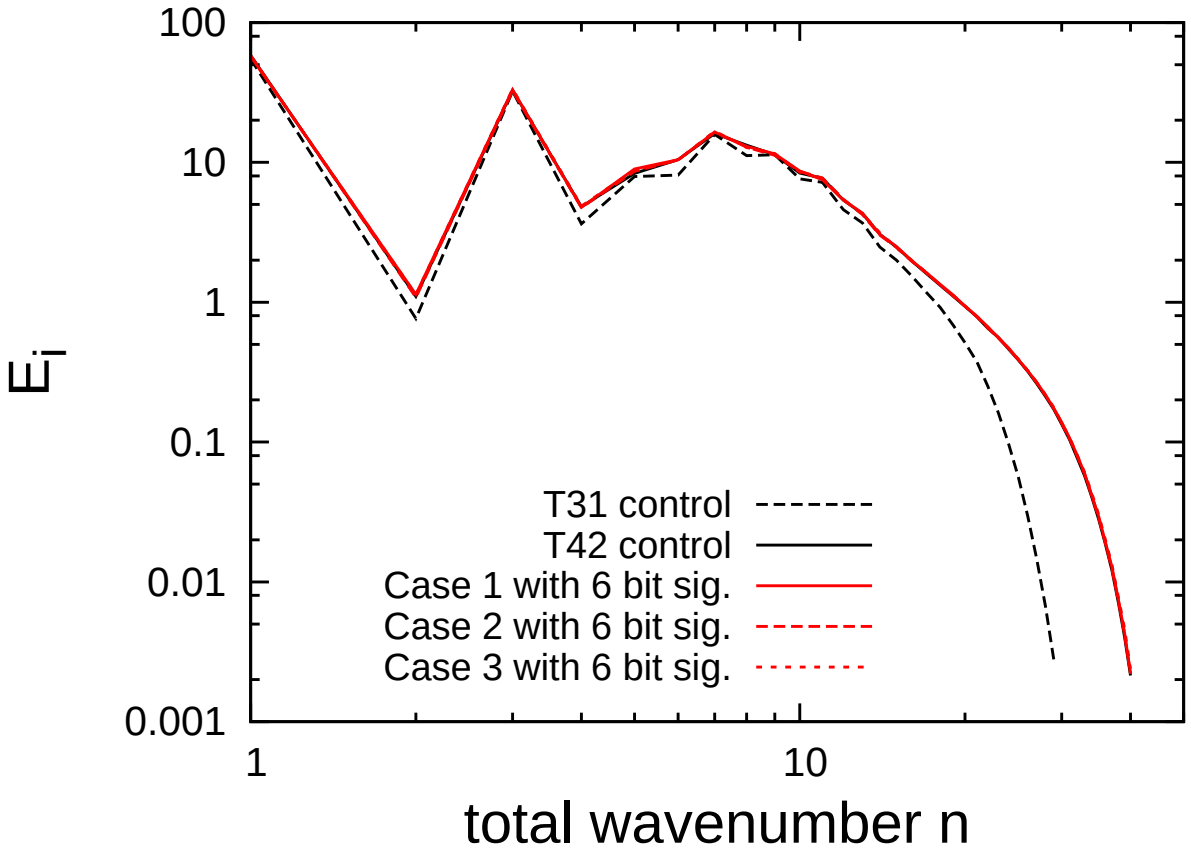


Promising!

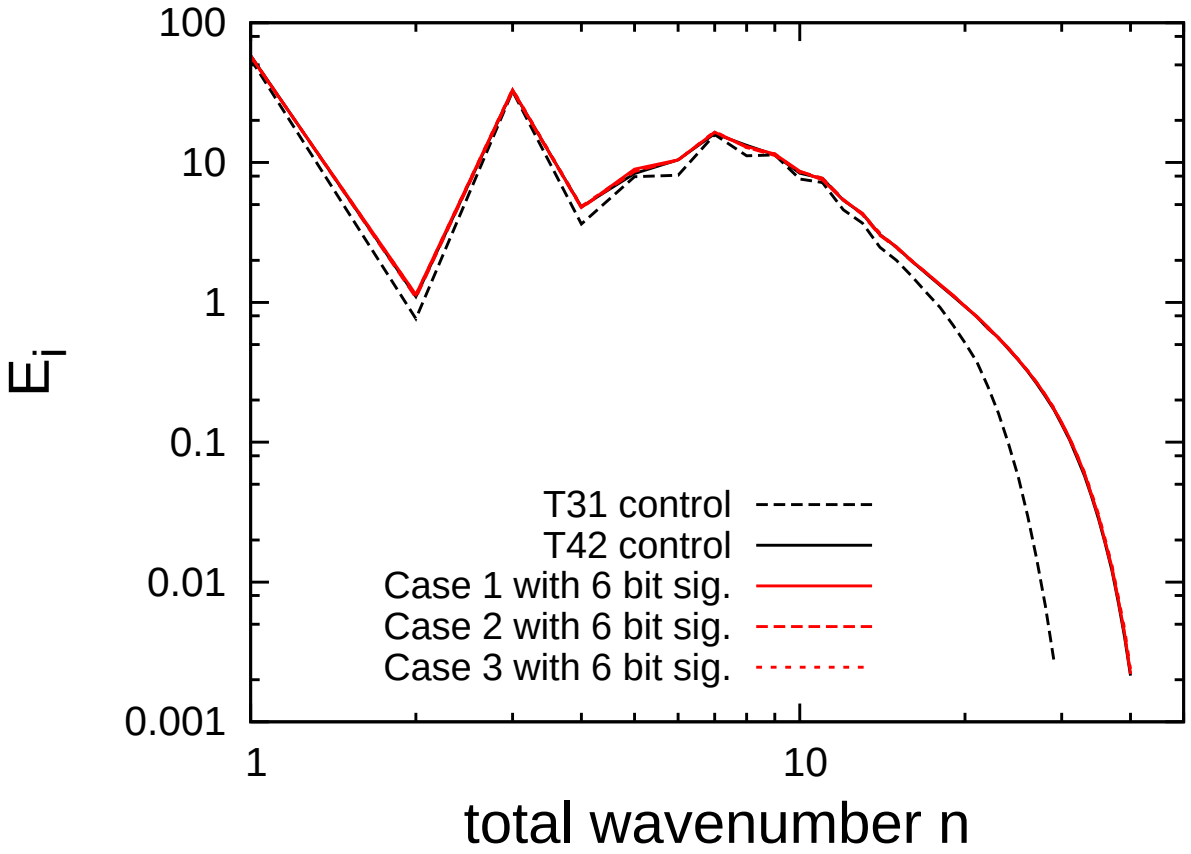
IGCM: Energy spectra



IGCM: Energy spectra



IGCM: Energy spectra



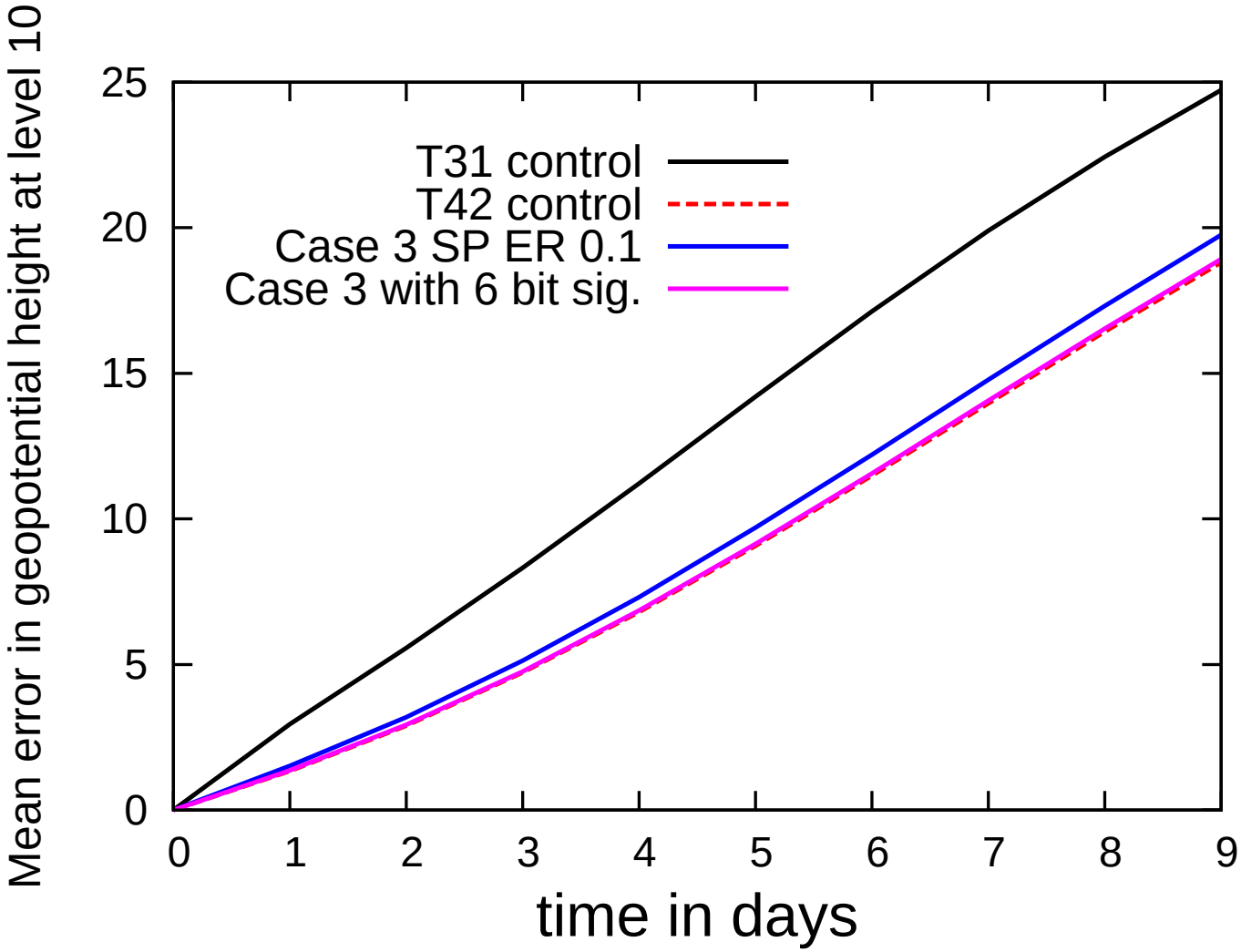
The energy spectra are the same.

Weather forecast with imprecise processing

Setup:

1. We calculate a truth by simulating the Held-Suarez test case with T84 resolution.
2. We initialise simulations with T31 and T42 resolution, using the truth.
3. We perform global forecasts with coarser resolution and compare with the truth.
4. We calculate the forecast error for geopotential height at the tenth vertical level (out of twenty).

Weather forecast with imprecise processing



Averaged over the globe and forty forecasts.

Conclusions

- The use of imprecise processing is promising high savings in energy demand, or a significant increase in performance.
- A crude implementation of imprecise processing will lead to problems.
- Scale separation is a necessary but sufficient tool to allow the use of imprecise processing in the dynamical core of a spectral model.
- We obtain promising results when emulating stochastic processors and very low floating point precision to calculate up to 84 % of the floating point operations.

Conclusions

- The use of imprecise processing is promising high savings in energy demand, or a significant increase in performance.
- A crude implementation of imprecise processing will lead to problems.
- Scale separation is a necessary but sufficient tool to allow the use of imprecise processing in the dynamical core of a spectral model.
- We obtain promising results when emulating stochastic processors and very low floating point precision to calculate up to 84 % of the floating point operations.

Shall we sacrifice precision for performance in weather and climate models?

Conclusions

- The use of imprecise processing is promising high savings in energy demand, or a significant increase in performance.
- A crude implementation of imprecise processing will lead to problems.
- Scale separation is a necessary but sufficient tool to allow the use of imprecise processing in the dynamical core of a spectral model.
- We obtain promising results when emulating stochastic processors and very low floating point precision to calculate up to 84 % of the floating point operations.

Shall we sacrifice precision for performance in weather and climate models?

Probably yes.

Outlook

- We perform higher resolution forecasts with IGCM using imprecise processing.
- We test if it is possible to use single precision for floating point numbers in large parts of the IFS.
- We test the use of imprecise processing in grid point models (SQG).
- We collaborate with the group around Krishna Palem to perform simulations of the atmosphere on real stochastic hardware.
- We want to test the use of data flow machines for weather and climate modelling.

Outlook

- We perform higher resolution forecasts with IGCM using imprecise processing.
- We test if it is possible to use single precision for floating point numbers in large parts of the IFS.
- We test the use of imprecise processing in grid point models (SQG).
- We collaborate with the group around Krishna Palem to perform simulations of the atmosphere on real stochastic hardware.
- We want to test the use of data flow machines for weather and climate modelling.

Thank you for your attention!