ED23C-0925

AGU 100 ADVANCING EARTH AND SPACE SCIENCE FALL MEETING
Washington, D.C. | 10–14 Dec 2018

GEOMAR

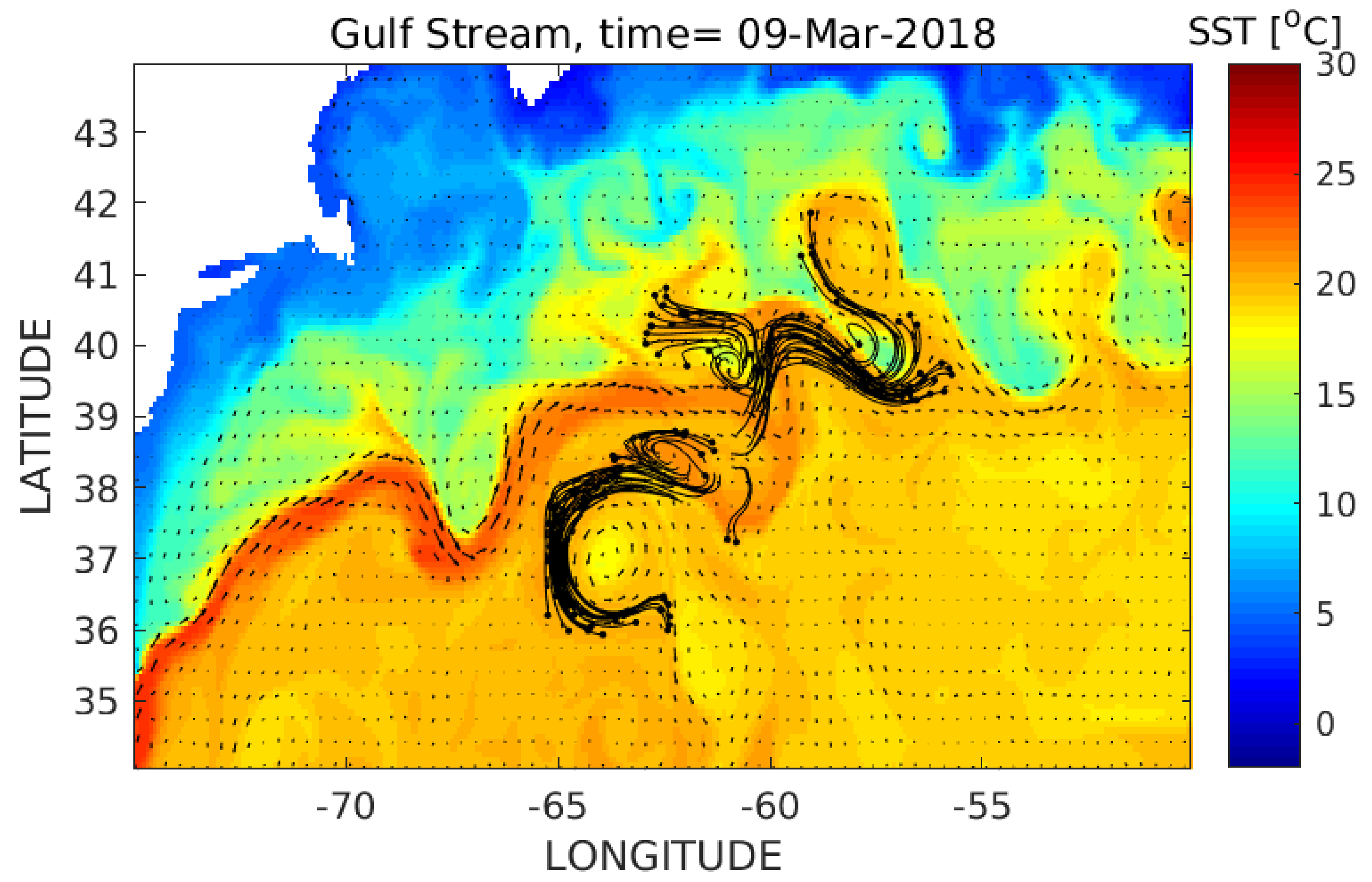
TEACHING STATISTICS WITH LAGRANGIAN TRAJECTORIES

Inga Monika Koszalka (ikoszalka@geomar.de)
GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany

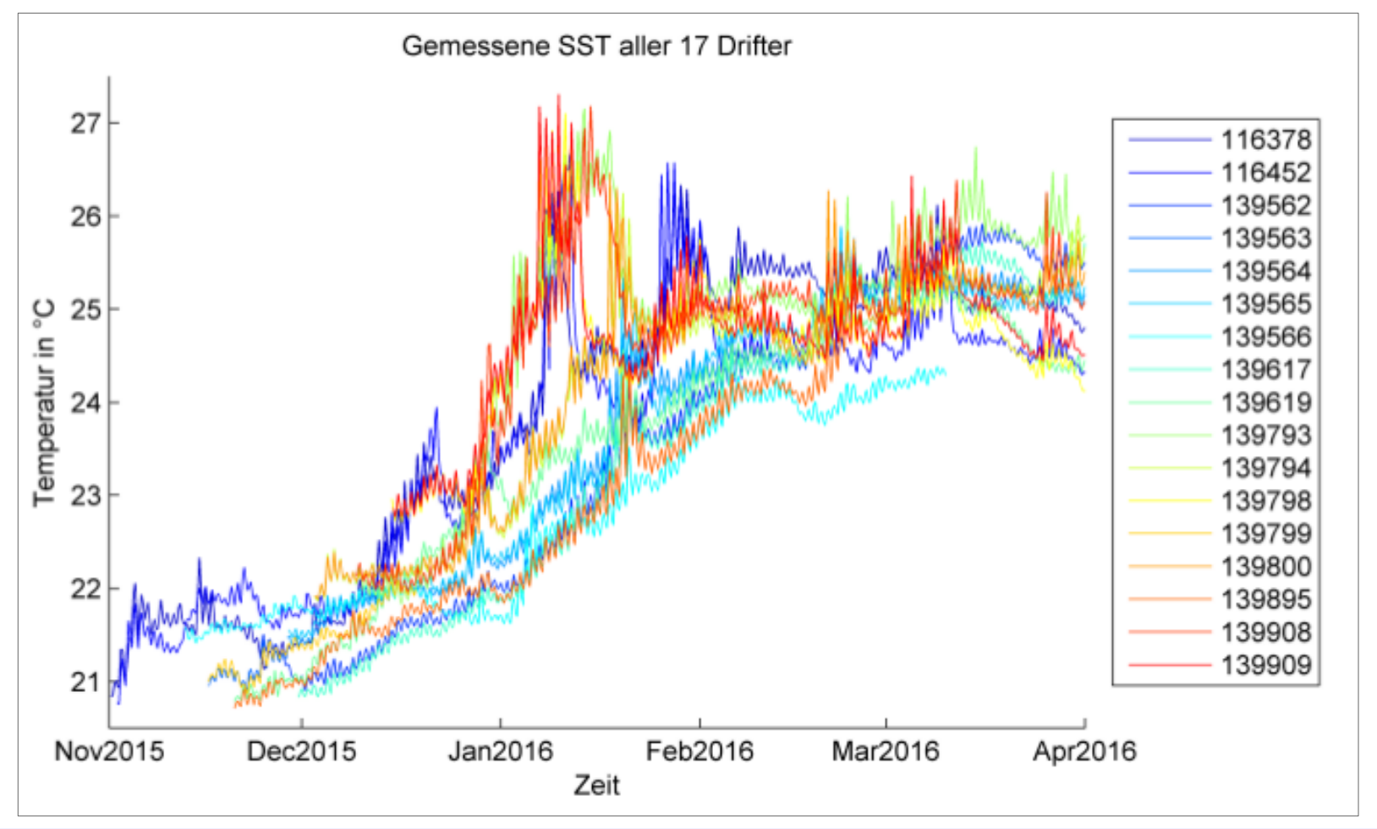
ABSTRACT

Have you ever felt your teaching of Statistics is dry, and also boring to students? Some topics are more challenging and grueling in this respect than others. Here, I am showing how certain statistical topics and concepts can be compellingly presented using Lagrangian (flow-following) surface drifter trajectories. It is thanks to the physical interpretation of Lagrangian trajectories, which encode information about the ocean flow kinematics and its spatio-temporal variability, that makes the process of applying statistical methods captivating.

→ Lagrangian trajectories provide appealing real-world material to make MsC and PhD students learn about caveats and conscious use of statistical tools.



Top: five-day long Lagrangian drifter trajectories in the Gulf-Stream. **Lagrangian trajectories** *refer to time series of positions of drifters floating in the ocean (real or simulated using output of ocean circulation models). Zonal and meridional velocity, temperature (below) and other properies can be measured along the trajectories. Note that drifters sample* **spatio-temporal variability**. *Under assumption of local homogeneity, Lagrangian trajectories and corresponding time series of water properties can be regarded as different realizations of the same* **stochastic (random) process**.



REFERENCES:

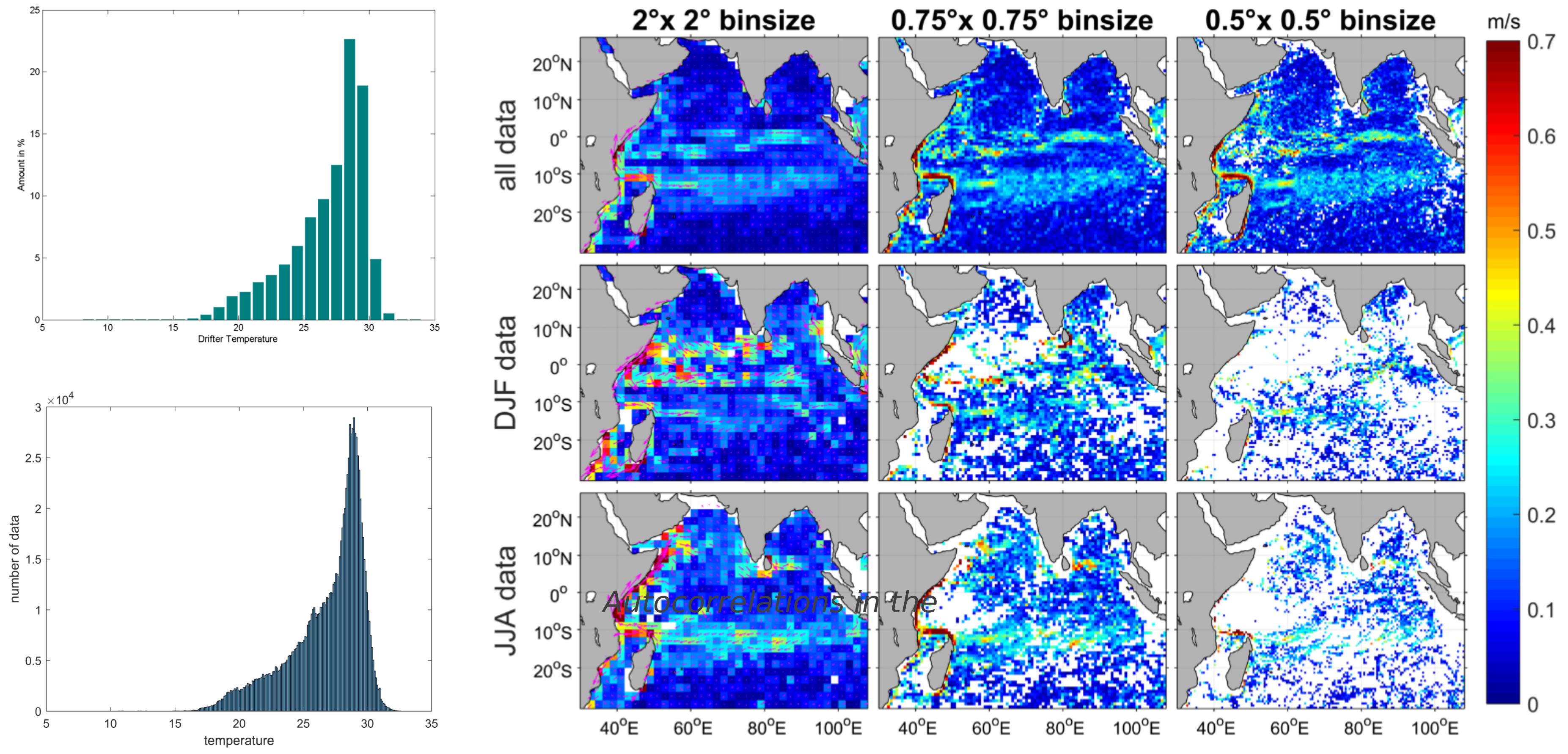
The material presented here stems from Master/PhD courses in “Data Analysis and Statistics” and “Lagrangian analysis and dispersion in the ocean” that I teach at GEOMAR/Kiel University, several figures are from student assignments.

PDF ESTIMATORS AND BANDWIDTH SELECTION

Probability distribution function is estimated by the histogram method which requires choosing a band (bin) size facing the trade-off between the bias and mean square error:

$$B(\widehat{f_X}(x)) \leq \frac{L}{K}, \quad \mathrm{Var}(\widehat{f_X}(x)) = K \cdot \frac{f_X(x^*)}{n} + \frac{{f_X}^2(x^*)}{n} \quad \mathrm{MSE}(\widehat{f_X}(x)) \leq \frac{L^2}{K^2} + K \cdot \frac{f_X(x^*)}{n} + \frac{{f_X}^2(x^*)}{n}$$

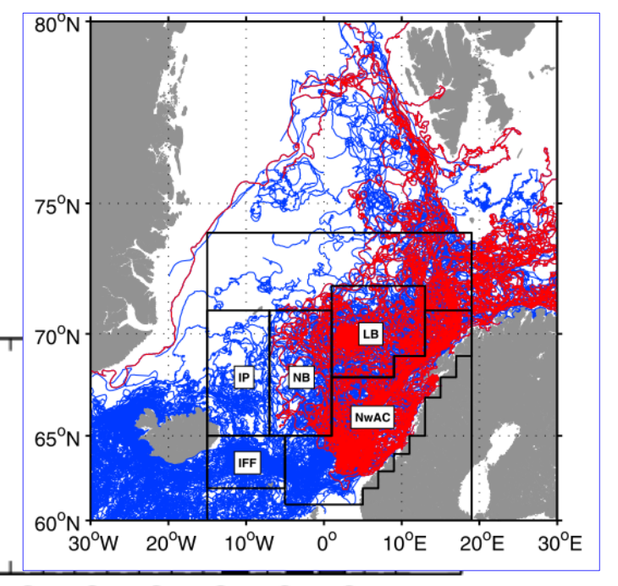
→ Students learn histogramization techniques and justify the chosen band width

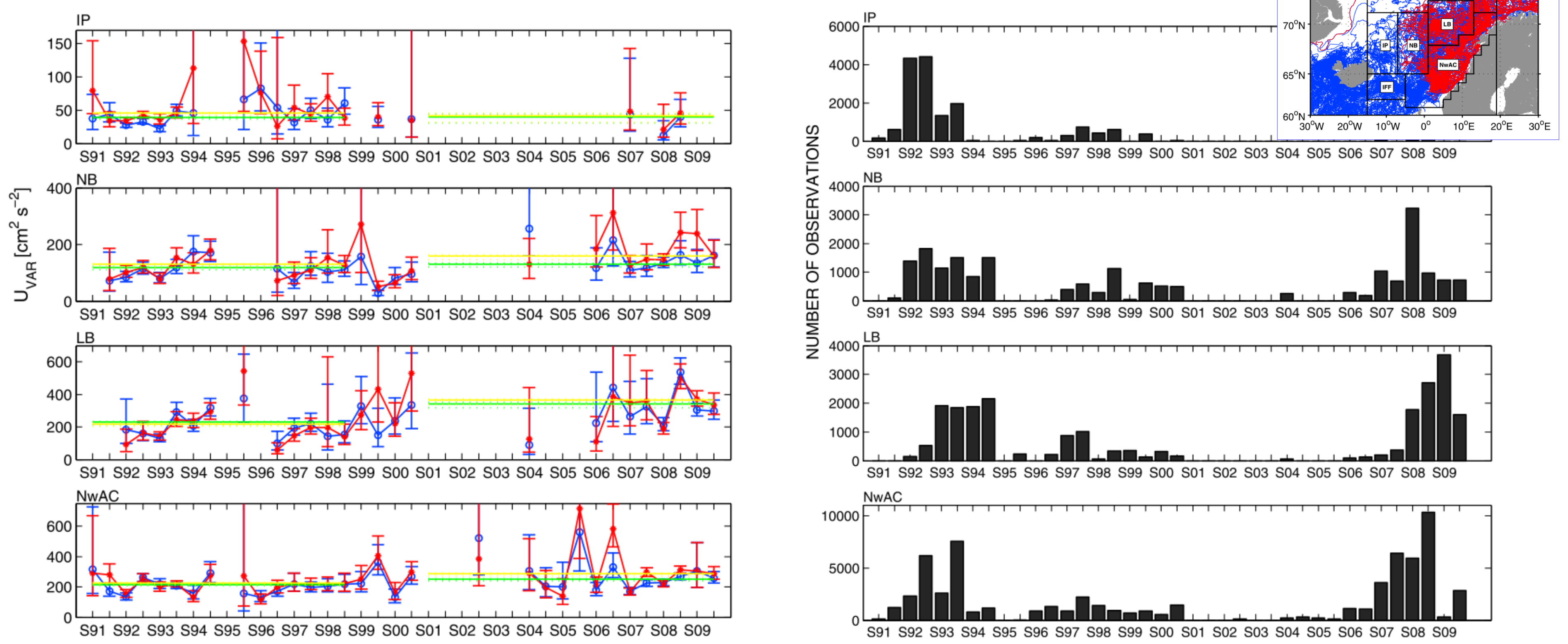


Left: Probability distributions of drifter temperatures in the Indian Ocean estimated by histograms with different band (bin) width. Right: Mean drifter speeds estimated by grouping Lagrangian observations into geographical bins (two-dimensional histograms). The bandwidth selection involves a trade-off between sufficient number of data for robust means (central limit theory) and the goal of resolving spatial variability of the current speed.

NUMBER OF DATA AND FALSE VARIABILITY

In this example, the mean Lagrangian velocities and variances (eddy kinetic energy) are calculated by grouping Lagrangian observations in larger regions.
Due to the dependence of variance on the number of observations,
false internannual variability can be concluded!



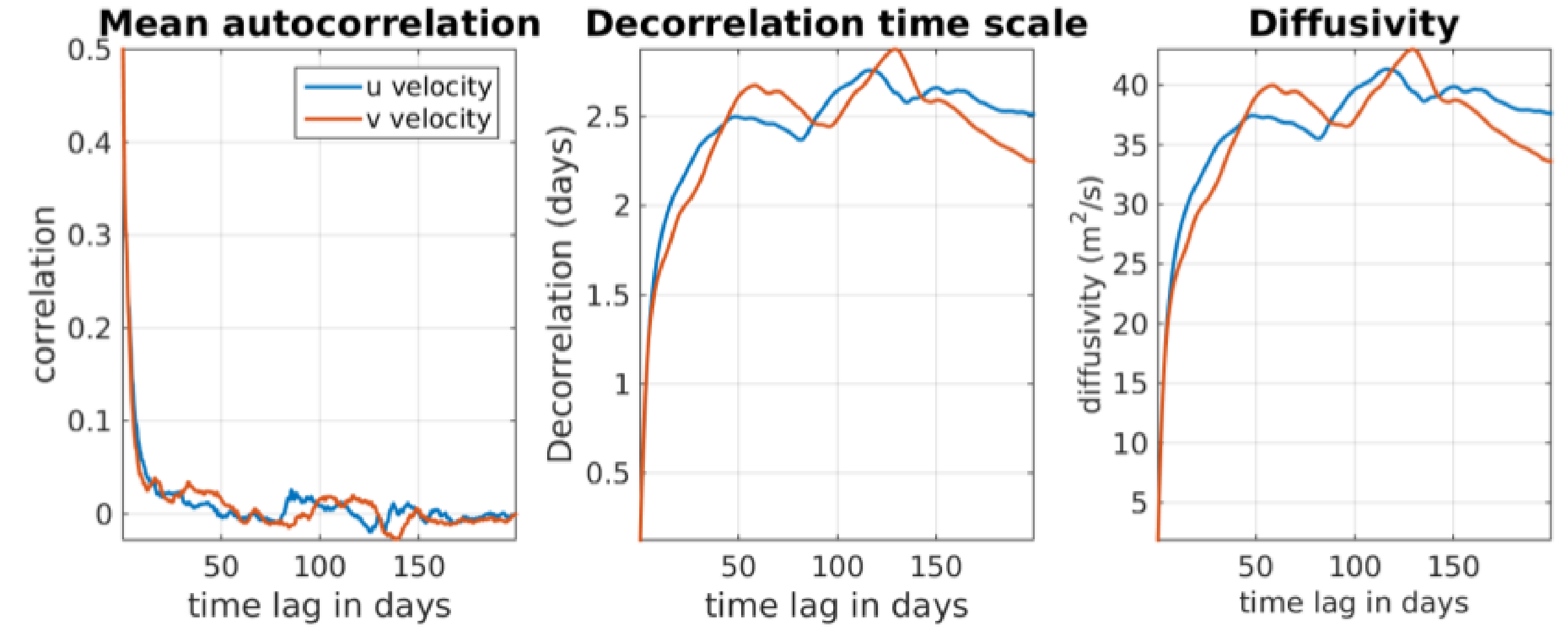


AUTOCORRELATION AND DEGREES OF FREEDOM

A Lagrangian time series can be viewed as realization from a stochastic process and, like other geophysical times series, are often auto-correlated. Estimation of **autocorrelation function** *(ACF):*

$$r_k = \frac{\sum_{i=1}^{N-k}(x_i - \overline{x})(x_{i+k} - \overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

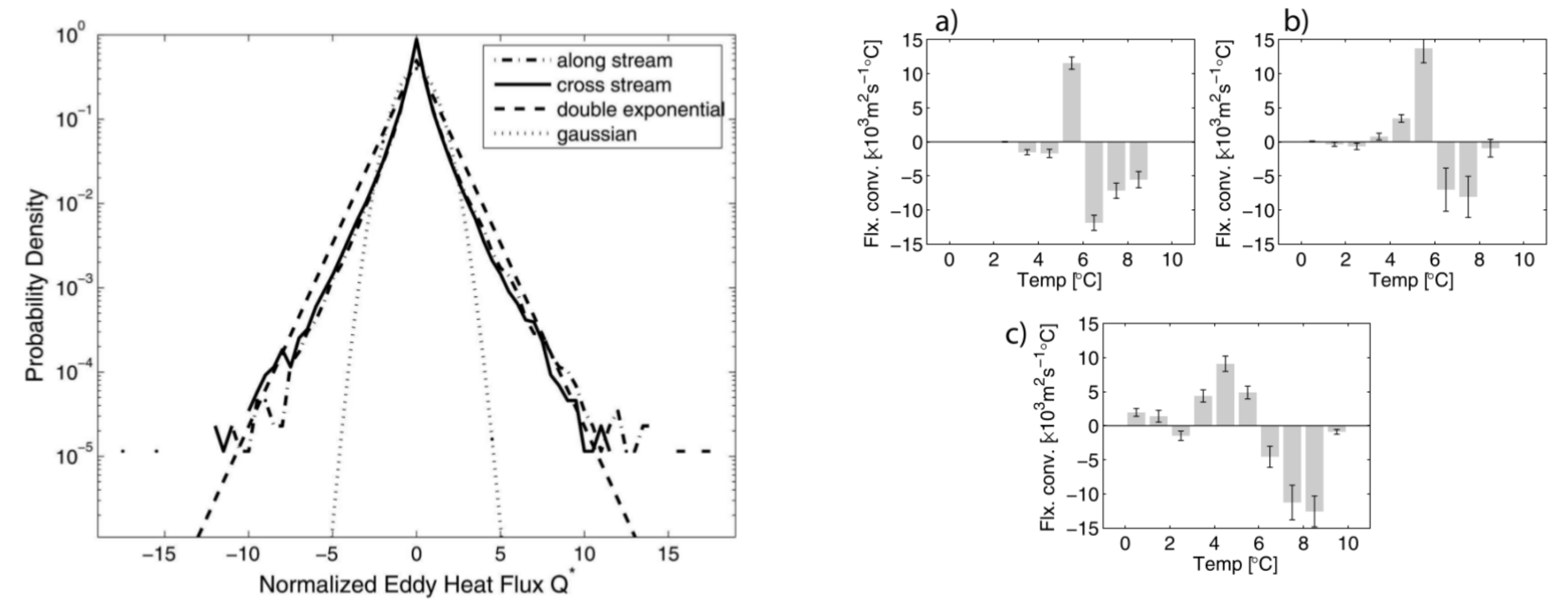
involves a choice between the unbiased estimator with higher MSE and the biased estimator with lower MSE. The **decorrelation time scale** *(Td) can be estimated from exponential decay (for AR(1)-like processes) or as integral time scale. Statistical inference methods (confidence intervals, hypothesis testing) require random variables to be iid. Inference for autocorrelated data thus require estimating of* **independent observations** *(Ni) and rescaling the degrees of freedom or shuffling or re-sampling the time series.*



→ Students learn techniques for ACF, Td and Ni estimation and their importance for statistical inference

CROSS-CORRELATIONS AND HEAT FLUXES

Autocorrelations in the data corrupt estimation of cross-correlation (false cross-correlations) and pose difficulty to estimate e.g., turbulent fluxes. These often require averaging over large (yet: homogeneous) regions or property classes to achieve robust statistics.



Left: PDFs of eddy heat fluxes from one timeseries. Right: Heat fluxes in temperature classes in the Nordic Seas.

HELMHOLTZ SPITZENFORSCHUNG FÜR GROSSE HERAUSFORDERUNGEN