# Visual Analytics for Large Scale Scientific Ensemble Datasets

## Fiscal Year 2018

Jonathan Woodring, Han-Wei Shen, and Tom Peterka

# Introduction

Scientific ensemble data sets have played increasingly more important roles for uncertainty quantification in various scientific and engineering domains, such as climate, weather, aerodynamics, and computational fluid dynamics. Ensembles are collections of data produced by simulations or experiments conducted with different initial conditions, parameterizations, or phenomenological models. They are usually used to describe complex systems, study sensitivities to initial conditions and parameters, and mitigate uncertainty. The goal of this proposal is to develop visual analytic techniques for large scale scientific ensemble data sets. Using ensemble simulations as an example, for a single run of such a simulation, there can be data generated in the range of several hundred gigabytes to tens of terabytes. A large scale ensemble dataset can consist of hundreds or thousands of such instances, with many variables in the form of scalar, vector, or tensor, and has a large number of samples in the high-dimensional input parameter space.

We proposed to research and develop methods for large-scale data analytics and visualization as applied to scientific data ensembles in several different topic areas: 1) *Exploration of Local Uncertainty with Distributions*, 2) *Exploration and Tracking of Ensemble Features*, and 3) *Exploration of Multivariate Ensemble Parameters.* Additionally, we proposed to tackle the scalability of these methods as applied towards DOE applications of interest: 1) *Automation of In Situ Ensemble Analytics* and 2) *Domain Specific and Laboratory Applications.* Below, we present selected results of our efforts in each of these aforementioned areas for FY 2018. Additionally for a primer, we also present a reference to our recent survey of ensemble data analytics.

## Visualization and Visual Analysis of Ensemble Data: A Survey [11]

Over the last decade, ensemble visualization has witnessed a significant development due to the wide availability of ensemble data, and the increasing visualization needs from a variety of disciplines. From the data analysis point of view, it can be observed that many ensemble visualization works focus on

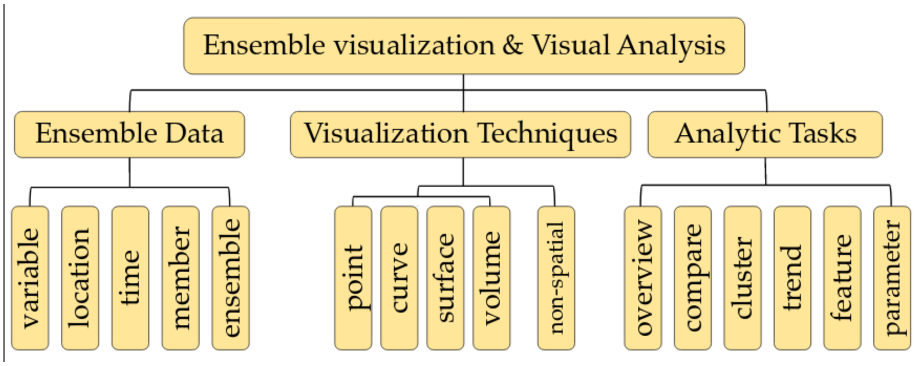


Figure 1: The ontology of ensemble visualization and analysis methods presented in our survey.

the same facet of ensemble data, use similar data aggregation or uncertainty modeling methods. However, the lack of reflections on those essential commonalities and a systematic overview of those works prevents visualization researchers from effectively identifying new or unsolved problems and planning for further developments.

In this paper, we take a holistic perspective and provide a survey of ensemble visualization. Specifically, we study ensemble visualization works in the recent decade, and categorize them from two perspectives: (1) their proposed visualization techniques; and (2) their involved analytic tasks. For the first perspective, we focus on elaborating how conventional visualization techniques (e.g., surface, volume visualization techniques) have been adapted to ensemble data; for the second perspective, we emphasize how analytic tasks (e.g., comparison, clustering) have been performed differently for ensemble data. From the study of ensemble visualization literature, we have also identified several research trends, as well as some future research opportunities.

Figure 1 shows the structure and breakdown of the different methods in our survey. We start with the fundamental concepts of ensemble data by answering questions like: what is ensemble data; how ensemble data is different from traditional scientific data; and what makes the visualization of ensemble data difficult. An intuitive data representation is formalized, which identifies the five orthogonal dimensions of ensemble data (i.e., variable, location, time, member, and ensemble). The representation is application independent, and thus, can be used to relate ensemble data from different disciplines. We additionally categorize the methods based on visualization tasks or analysis tasks.

# Exploration of Local Uncertainty with Distributions

## Uncertainty Visualization Using Copula-Based Analysis in Mixed Distribution Models [7]

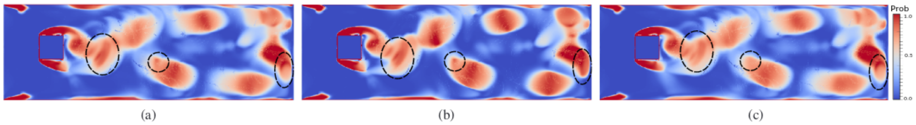


Figure 2: An example of using copula based modeling vs. Gaussian only or KDE-only methods

Distributions are often used to model uncertainty in many scientific datasets. To preserve the correlation among the spatially sampled grid locations in the dataset, various standard multivariate distribution models have been proposed in visualization literature. These models treat each grid location as a univariate random variable which models the uncertainty at that location. Standard multivariate distributions (both parametric and nonparametric) assume that all the univariate marginals are of the same type/family of distribution. But in reality, different grid locations show different statistical behavior which may not be modeled best by the same type of distribution. In this paper, we propose a new multivariate uncertainty modeling strategy to address the needs of uncertainty modeling in scientific datasets.

Our proposed method is based on a statistically sound multivariate technique called *Copula*, which makes it possible to separate the process of estimating the univariate marginals and the process of modeling dependency, unlike the standard multivariate distributions. The modeling flexibility offered by our proposed method makes it possible to design distribution fields which can have different types of distribution (Gaussian, Histogram, KDE etc.) at the grid locations, while maintaining the correlation structure at the same time. Depending on the results of various standard statistical tests, we can choose an optimal distribution representation at each location, resulting in a more cost efficient modeling without significantly sacrificing on the analysis quality. To demonstrate the efficacy of our proposed modeling strategy, we extract and visualize uncertain features like isocontours and vortices in various real world datasets. We also study various modeling criterion to help users in the task of univariate model selection.

In Figure 2, we show an example of modeling an ensemble of vortex core simulations using our *Copula* method, compared with single model-type strategies. The marked regions in the images highlight the differences in modeling of the probabilities of vortex cores appearing, where (a) is the *Coupla* method, (b) is a Gaussian-only method, and (c) is a Kernel Density Estimate (KDE) only method. The result generated by our proposed method as shown by (a) is a mixed representation of both the types of distributions. Therefore, the regions with high certainty of following a Gaussian distribution are able to show the probable vortex structures that (b) reflects. Whereas, regions where we used KDE to model the data where able to show results similar to (c). For example, the region marked by the right-most black circle highlights a feature which was missed out by assuming Gaussian distribution, but was captured by both KDE and our mixed representation.

# Exploration and Tracking of Ensemble Features

## eFESTA: Ensemble Feature Exploration with Surface [4]

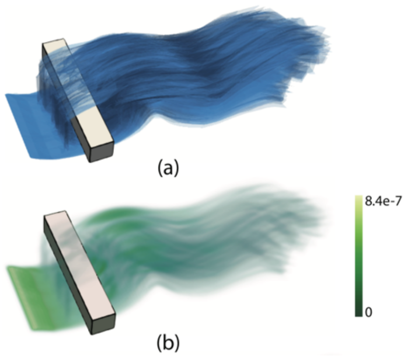


Figure 3: An example of surface density estimates.

Ensemble simulations are becoming prevalent in various scientific and engineering domains, such as climate, weather, aerodynamics, and computational fluid dynamics. An ensemble is a collection of data produced by simulations for the same physical phenomenon conducted with different initial conditions, parameterizations, or phenomenological models. Ensemble simulations are used to simulate complex systems, study sensitivities to initial conditions and parameters, and mitigate uncertainty. For example, in numerical weather prediction, ensemble forecasts with different fore- cast models and initial conditions are widely used to indicate the range of possible future states of the atmosphere.

We propose surface density estimate (SDE) to model the spatial distribution of surface features—isosurfaces, ridge surfaces, and streamsurfaces—in 3D ensemble simulation data. The inputs of SDE computation are surface features represented as polygon meshes, and no field datasets are required (e.g., scalar fields or vector fields). The SDE is defined as the kernel density estimate of the infinite set of points on the input surfaces and is approximated by accumulating the surface densities of triangular patches. We also propose an algorithm to guide the selection of a proper kernel bandwidth for SDE computation. An ensemble Feature Exploration method based on Surface densiTy EstimAtes (eFESTA) is then proposed to extract and visualize the major trends of ensemble surface features. For an ensemble of surface features, each surface is first transformed into a density field based on its contribution to the SDE, and the resulting density fields are organized into a hierarchical representation based on the pairwise distances between them. The hierarchical representation is then used to guide visual exploration of the density fields as well as the underlying surface features. We demonstrate the application of our method using isosurface in ensemble scalar fields, Lagrangian coherent structures in uncertain unsteady flows, and streamsurfaces in ensemble fluid flows.

As shown in Figure 3, the input surfaces and the output density field are visualized in (a) and (b), respectively. Given an ensemble of surfaces, a straightforward density estimation approach is to define a regular grid over the surfaces, and then count the number of surfaces intersecting each grid cell. However, after discretizing the surfaces with respect to a given grid, the information of the surface patches (e.g., location, orientation, and shape) within each grid cell is lost, which introduces discretization error into the density estimation results. Although increasing the grid resolution can reduce the discretization error, the computation cost increases. In this work, we propose SDE, which generalizes the kernel density estimate (KDE) from discrete sample points to the infinite set of points on input surfaces. We approximate SDE of the input surfaces by accumulating the surface densities of triangular patches, which can be calculated based on bivariate normal integrals with efficient GPU computation.

# Exploration of Multivariate Ensemble Parameters

## Extreme-Scale Stochastic Particle Tracing for Uncertain Unsteady Flow Visualization and Analysis [2]

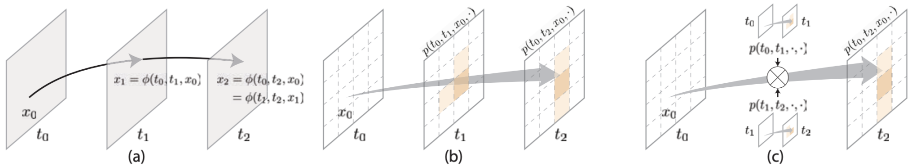

$x_0$
$t_0$
$x_1 = \phi(t_0, t_1, x_0)$
$t_1$
$x_2 = \phi(t_0, t_2, x_0)$
$= \phi(t_1, t_2, x_1)$
$t_2$
(a)
$x_0$
$t_0$
$p(t_0, t_1, x_0, \cdot)$
$t_1$
$p(t_0, t_2, x_0, \cdot)$
$t_2$
(b)
$x_0$
$t_0$
$t_0$
$t_1$
$p(t_0, t_1, \cdot, \cdot)$
$p(t_1, t_2, \cdot, \cdot)$
$t_1$
$t_2$
$p(t_0, t_2, x_0, \cdot)$
$t_2$
(c)

Figure 4: An example of stochastic flow maps (SFM).

Visualizing and analyzing data with uncertainty are important in many science and engineering domains, such as computational fluid dynamics, climate, weather, and materials sciences. Instead of analyzing deterministic data resulted from statistical aggregation, scientists can gain more understanding by investigating uncertain data that are derived and quantified from experiments, interpolation, or numerical ensemble simulations. For example, typical analyses of uncertain flows involve finding possible pollution diffusion paths in environmental sciences with uncertain source-destination queries and locating uncertain flow boundaries in computational fluid dynamics models with uncertain Lagrangian analysis.

We present an efficient and scalable solution to estimate uncertain transport behaviors—stochastic flow maps (SFMs)—for visualizing and analyzing uncertain unsteady flows. Computing flow maps from uncertain flow fields is extremely expensive because it requires many Monte Carlo runs to trace densely seeded particles in the flow. We reduce the computational cost by decoupling the time dependencies in SFMs so that we can process shorter sub time intervals independently and then compose them together for longer time periods. Adaptive refinement is also used to reduce the number of runs for each location. We parallelize over tasks—packets of particles in our design—to achieve high efficiency in MPI/thread hybrid programming. Such a task model also enables CPU/GPU coprocessing. We show the scalability on two supercomputers, Mira (up to 256K Blue Gene/Q cores) and Titan (up to 128K Opteron cores and 8K GPUs), that can trace billions of particles in seconds.

As shown in Figure 4, the key to achieve parallelism is to decouple sub time intervals to remove time dependencies. The SFM can be estimated by composing the intermediate results from each subinterval. We also derived the theoretical error bound of decoupled SFM estimate, which is related to the number of subintervals, the mesh discretization, and the smoothness of the SFM distribution.

## Automation of In Situ Ensemble Analytics

### Parallel Partial Reduction for Large-Scale Data Analysis and Visualization [3]

We present a novel partial reduction algorithm to aggregate sparsely distributed intermediate results that are generated by data-parallel analysis and visualization algorithms. Applications of partial reduction include flow trajectory analysis, big data online analytical processing, and volume rendering. Unlike traditional full parallel reduction that exchanges dense data across all processes, the purpose of partial reduction is to exchange only intermediate results that correspond to the same query, such as line segments of the same flow trajectory.

To this end, we design a three-stage algorithm that minimizes the communication cost: (1) partitioning the result space into groups; (2) constructing and optimizing the reduction partners for each group; and (3) initiating collective reduction operations for all groups concurrently. Both theoretical and empirical

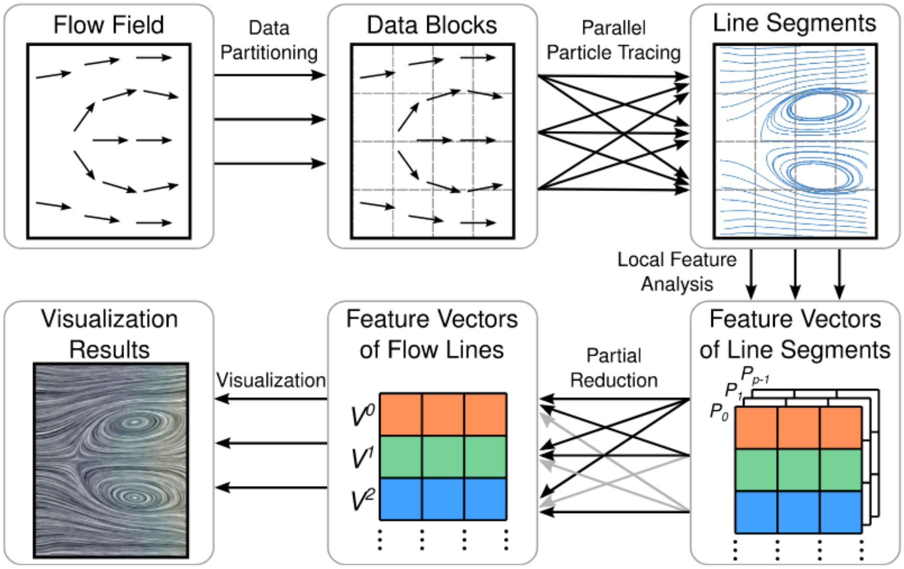
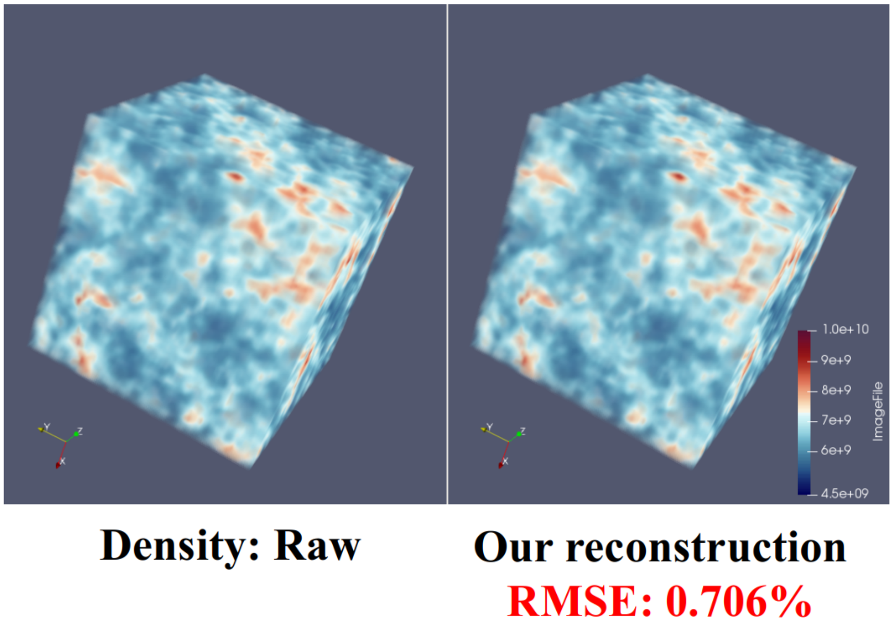


Figure 5: An example workflow of Lagrangian-based flow analyses with parallel partial reduction

analyses show that our algorithm outperforms the traditional methods when the intermediate results are sparsely distributed. We also demonstrate the effectiveness of our algorithm for flow visualization, big log data analysis, and volume rendering.

Figure 5 shows an application of Lagrangian-based flow visualization applications that can benefit from our partial reduction methods. For example, the total length of flow lines can be calculated by summing the lengths of flow line segments; the operator for the reduction is "add." Importantly for ensemble statistical-based analyses, streamline statistics (histograms, mean, and variance) are also based on the add operator. The line integral convolutions (LIC) result of a given streamline seed can be computed by summing the convolution values across processes. The reduction of flow line queries, such as predicates and pattern matching, is based on the logical "or" operator.



Density: Raw

Our reconstruction

RMSE: 0.706%

Figure 6: A reconstruction of Nyx simulation data using statistical super resolution compared to the original data.

# Domain Specific and Laboratory Applications

## Cosmology

Cosmological models have been developed to simulate the large-scale evolution of the structure of the universe. Though, these are parameterized by initial conditions, consisting of many initial physical parameters, and the exact values of the initial conditions are not known. Physicists, such as scientists at LBL with the Nyx simulation, are searching for these initial conditions of our universe. To do this, they study ensembles of the simulated distributions of matter as compared to today's observed universe. However, the number of physical parameters of interest create a huge data space to search and analyze from high-resolution cosmological simulations. Assessing such massive datasets in post-analysis will be slow and hinder the time to results due to the limited I/O bandwidth and storage capacity. Developing techniques to reduce the data size in situ, meet the I/O bandwidth and disk storage constraints, and provide the desired scientific accuracy is critical for cosmology.

In response, we evaluated two promising techniques from previous research applied towards DOE applications: "Incremental GMM-based (Gaussian Mixture model) Emulator" and "Statistical-based Super-resolution," to achieve our goal. The GMM-based Emulator uses multi-variate GMMs to compactly summarize the data from initial simulation parameters and quantities of interest. The data set from a given initial condition can then be reconstructed from the emulator in the post-analysis stage. The GMM-based emulator, itself, is built incrementally, over time, when the data from a new simulation is generated in the parameter space. This allows for both smaller storage footprint and faster simulation and analysis time due to reduced data.

Secondly, Statistical-based Super-resolution statistically down-samples cosmological data, in situ, and later reconstructs the smaller, down-sampled data for analysis, to full resolution. An example of the original data compared to our reconstruction is shown in Figure 6. The process involves a one-time task that collects a small subset of full-resolution data generated from the ensemble. This creates a prior knowledge data set for super-resolution reconstruction. Then during later simulations, statistical-based down-sampling is applied, reduces the size of data saved. Combined with the down-sampled data, the prior knowledge data set is used to compensate for the lack of spatial information and details to recover it to full-resolution and high accuracy. This work has been submitted for peer-review to Pacific Visualization 2019.

# Citations

[1] Soumya Dutta, Han-Wei Shen, and Jen-Ping Chen. *In Situ Prediction Driven Feature Analysis in Jet Engine Simulations.* Pacific Visualization Symposium (PacificVis), 2018 IEEE, 66-75.

[2] Hanqi Guo, Wenbin He, Sangmin Seo, Han-Wei Shen, Emil Mihai Constantinescu, Chunhui Liu, and Tom Peterka, *Extreme-Scale Stochastic Particle Tracing for Uncertain Unsteady Flow Visualization and Analysis.* IEEE Transactions on Visualization and Computer Graphics, 2019.

[3] Wenbin He, Hanqi Guo, Tom Peterka, Sheng Di, Franck Cappello, and Han-Wei Shen. *Parallel Partial Reduction for Large-Scale Data Analysis and Visualization.* In Proceedings of 2018 IEEE Symposium on Large Data Analysis and Visualization, 2018.

[4] Wenbin He, Hanqi Guo, Han-Wei Shen, and Tom Peterka, *eFESTA: Ensemble Feature Exploration with Surface Density Estimates.* IEEE Transactions on Visualization and Computer Graphics, 2019.

[5] Subhashis Hazarika, Ayan Biswas, Soumya Dutta, and Han-Wei Shen. *Information Guided Exploration of Scalar Values and Isocontours in Ensemble Datasets.* Entropy 2018, 20(7), 540.

[6] Subhashis Hazarika, Soumya Dutta, Han-Wei Shen, Jen-Peng Chen. *CoDDA: A Flexible Copula-based Distribution Driven Analysis Framework for Large-Scale Multivariate Data.* IEEE Transactions on Visualization and Computer Graphics.

[7] Subhashis Hazarika, Ayan Biswas, Han-Wei Shen. *Uncertainty Visualization Using Copula-Based Analysis in Mixed Distribution Models.* IEEE Transactions on Visualization and Computer Graphics, 24(1): 934-943 (2018).

[8] Cheng Li, Joachim Moortgat, and Han-Wei Shen. *An Automatic Data Deformation Approach for Occlusion Free Egocentric Data Exploration.* Pacific Visualization Symposium (PacificVis), 2018 IEEE, 215-224.

[9] Junpeng Wang, Liang Gou, Han-Wei Shen, Hao Yang. *DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks.* IEEE Transactions on Visualization and Computer Graphics.

[10] Junpeng Wang, Liang Gou, Hao Yang, and Han-Wei Shen. *GANViz: A Visual Analytics Approach to Understand the Adversarial Game.* IEEE Transactions on Visualization and Computer Graphics, 24 (6), 1905-1917 (2018).

[11] Junpeng Wang, Subhashis Hazarika, Cheng Li, Han-Wei Shen. *Visualization and Visual Analysis of Ensemble Data: A Survey.* IEEE Transactions on Visualization and Computer Graphics, 2018.

[12] Ko-Chih Wang, Naeem Shareef, and Han-Wei Shen. *Image and Distribution Based Volume Rendering for Large Data Sets.* Pacific Visualization Symposium

(PacificVis), 2018 IEEE, 26-35.

[13] Tzu-Hsuan Wei, Soumya Dutta, and Han-Wei Shen. *Information Guided Data Sampling and Recovery using Bitmap Indexing.* Pacific Visualization Symposium (PacificVis), 2018 IEEE, 56-65.

# Talks

- **Dagstuhl, Germany** – Seminar *In Situ Scientific Data Visualization* – July 1, 2018
- **ChinaVis 2008** – Tutorial *Information Theory for Data Analysis and Visualization* – July 7, 2018
- **Peking University** – Class Lecture *Distributed Based Data Modeling, Analysis, and Visualization* – July 22, 2018

# Awards

- **Best Paper** – Junpeng Wang, Liang Gou, Hao Yang, and Han-Wei Shen. *GANViz: A Visual Analytics Approach to Understand the Adversarial Game.* IEEE Transactions on Visualization and Computer Graphics, 24 (6), 1905-1917 (2018). *IEEE PacificVis 2018*

- **Best Paper, Honorable Mention** – Wenbin He, Hanqi Guo, Tom Peterka, Sheng Di, Franck Cappello, and Han-Wei Shen. *Parallel Partial Reduction for Large-Scale Data Analysis and Visualization.* In Proceedings of 2018 IEEE Symposium on Large Data Analysis and Visualization, 2018. *LDAV 2018*

- **Best Paper, Honorable Mention** – Junpeng Wang, Liang Gou, Han-Wei Shen, Hao Yang. *DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks.* IEEE Transactions on Visualization and Computer Graphics. *IEEE VAST 2018*

# Supported Students

- Hazarika, Subhashis (at Ohio State)
- He, Wenbin (at Ohio State)
- Wang, Ko-Chih (at LANL)
- Xu, Jiayi (at LANL)