# FlowBench : A Large Scale Benchmark for Flow Simulation over Complex Geometries (Supplementary Material)

**Ronak Tali[1]*, Ali Rabeh[1]*, Cheng-Hau Yang[1]*, Mehdi Shadkhah[1], Samundra Karki[1] Abhisek Upadhyaya[2], Suriya Dhakshinamoorthy[1], Marjan Saadati[1], Soumik Sarkar[1], Adarsh Krishnamurthy[1], Chinmay Hegde[2], Aditya Balu[1], Baskar Ganapathysubramanian[1]**

**[1]Iowa State University**
{rtali,arabeh,chenghau,mehdish, samundra,snarayan,marjansd,soumiks, adarsh,baditya,baskarg}@iastate.edu

**[2]New York University**
{au2216,chinmay.h}@nyu.edu

## 1 Description of Data

**Table 1:** *Formulaic description of the input and output tensors. 3000/6000/1150/500 are sample sizes for the dataset. 240 is the number of equi-spaced time snapshots for the FPO case; $x, y(, z)$ are the dimensions of a field. E.g., $Y[0, 1, :, :]$ indicates the pointwise $v$ velocity over the entire grid. We denote by $C$ - a single channel split into equal halves containing the coefficient of drag and coefficient of lift. We denote by $C^*$ - a single channel split into two equal fourths and a half containing the coefficient of drag, coefficient of lift, and the Nusselt number, respectively.*

| Dataset | Dim. | Input Tensor | Output Tensor |
|---|---|---|---|
| LDC - NS | 2 | $X[3000][Re, g, s][x][y]$ | $Y[3000][u, v, p, C][x][y]$ |
| LDC - NS+HT | 2 | $X[5990][Re, Gr, g, s][x][y]$ | $Y[5990][u, v, p, \theta, C^*][x][y]$ |
| FPO - NS | 2 | $X[1150][Re, g, s][x][y]$ | $Y[1150][240][u, v, p][x][y]$ |
| LDC - NS | 3 | $X[500][Re, g, s][x][y][z]$ | $Y[500][u, v, w, p][x][y][z]$ |

## 2 Datasheet for our Dataset

We follow the datasheet proposed in [1] for documenting our FlowBench dataset.

1. Motivation
   (a) For what purpose was the dataset created?
   The dataset was created to enable comprehensive evaluation of data-driven models that predict complex 2D/3D flow physics around a range of geometrical objects in steady-state and transient situations.
   (b) Who created the dataset and on behalf of which entity?
   The dataset was created by the Baskar Ganapathysubramanian Group (ComPM Lab) at Iowa State University, Ames, IA.
   (c) Who funded the creation of the dataset? FlowBench was funded in part by the the AI Research Institutes program supported by USDA-NIFA under AI Institute: for Resilient Agriculture, Award No. 2021-67021-35329, NSF under awards 1954556, 2323716, and 2053760.

*These authors contributed equally to this work.

Preprint. Under review.

(d) Any other Comments?
**[NA]**

2. Composition

(a) What do the instances that comprise the dataset represent?
Each instance captures some cardinal flow variables, such as velocities, pressure, or temperature defined over the entire domain of interest. In addition, each such field is packaged together as a single tensor, the details of which are discussed in Details in Table 1.

(b) How many instances are there in total?
10,650

(c) Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
We sample across geometries and flow operating conditions. Geometries are further classified into parametric and non-parametric families. Flow operating conditions are characterized by non-dimensional numbers (Reynolds and Grashof). Our dataset includes a dense, random collection of flow simulations sampled from this space of geometry $\times$ operating conditions. However, given that the space is a continuous space, it cannot contain all possible instances. For this reason, we have provided the complete code to the end user to generate more random geometries and, by implication, more input data instances.

(d) What data does each instance consist of?
Each data is a simulation outcome.
The simulation inputs are the shape of the object, and the operating conditions. The object's shape is provided in two (field) formats: binary mask and signed distance field (SDF). The operating conditions – Reynolds number, $Re$, and Grashof number $Gr$ – are provided as concordant fields. If $Gr$ is not provided, that simulation was performed for $Gr = 0$.
The simulation output consists of (a) fluid velocity in the domain given as a 2- or 3-dimensional field $(u, v, w)$, (b) pressure in the domain given as a field, (c) temperature (if flow thermal simulation case) in the domain given as a field, (d) Coefficient of lift, coefficient of drag, and Nusselt number as concordant fields
In Table 1, we have presented a detailed tensor formula that completely explains the constitution of each sample instance present in our dataset.

(e) Is there a label or target associated with each instance?
See the previous point.

(f) Is any information missing from individual instances?
**[No]**

(g) Are relationships between individual instances made explicit?
**[Yes]** . Details in Table 1.

(h) Are there recommended data splits?
**[No]**

(i) Are there any errors, sources of noise, or redundancies in the dataset?
**[No]**

(j) Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
**[Yes]** . Our dataset is self-contained.

(k) Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?
**[No]**

(l) Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
**[No]**

(m) Does the dataset relate to people?
**[No]**

(n) Does the dataset identify any subpopulations (e.g., by age, gender)?
**[No]**

(o) Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
**[No]**

(p) Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?
**[No]**

(q) Any other Comments?
**[No]**

3. Collection Process

(a) How was the data associated with each instance acquired?
Given a set of flow conditions and geometries, detailed flow simulations (using an in house validated simulator) were run and monitored to ensure results were properly convergent.

(b) What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
Flow simulations – specifically finite element simulation – were used to generate the data using our in-house simulation framework [2].

(c) If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
See Point 2(c). The flow conditions are randomly generated, using a Sobel sequence generator.

(d) Who was involved in the data collection process (e.g., students, crowdworkers, contractors), and how were they compensated (e.g., how much were crowdworkers paid)?
Only the authors of the paper were involved. No third parties or contract workers were involved.

(e) Over what timeframe was the data collected?
May 2024 to June 2024.

(f) Were any ethical review processes conducted (e.g., by an institutional review board)?
**[NA]**

(g) Does the dataset relate to people?
**[No]**

(h) Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
**[NA]**

(i) Were the individuals in question notified about the data collection?
**[NA]**

(j) Did the individuals in question consent to the collection and use of their data?
**[NA]**

(k) If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
**[NA]**

(l) Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?
**[NA]**

(m) Any other Comments?
**[No]**

4. **Preprocessing, Cleaning and Labeling**

(a) Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
All our raw data was generated using our in-house FEM software. The raw data was (sub)sampled and cropped using the ParaView tool [3]. These processed fields were packaged as numpy tensors for immediate use in Deep Learning frameworks.

(b) Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? **[Yes]**

(c) Is the software used to preprocess/clean/label the instances available?
We used either the ParaView tool [3] or in-house scripts - released in the "Code" section of our website

(d) Any other Comments?
**[No]**

5. Uses

(a) Has the dataset been used for any tasks already?
**[Yes]** . We have used the dataset to benchmark three leading Neural Operator models in the manuscripts.

(b) Is there a repository that links to any or all papers or systems that use the dataset?
**[No]**

(c) What (other) tasks could the dataset be used for?
**[NA]**

(d) Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
**[No]**

(e) Are there tasks for which the dataset should not be used?
**[No]**

(f) Any other Comments?
**[No]**

6. Distribution

(a) Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?
**[Yes]**

(b) How will the dataset be distributed (e.g., tarball on website, API, GitHub)?
We use an institutional data distributional service provided by Iowa State University, Ames, IA. This service allows direct download of our datasets using a URL, which has been included in our main paper.

(c) When will the dataset be distributed?
Upon acceptance of the paper.

(d) Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
Licensed under CC BY-NC 4.

(e) Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
**[NA]** . Third parties are not involved.

(f) Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?
**[NA]**

(g) Any other Comments?
**[No]**

7. Maintenance

(a) Who is supporting/hosting/maintaining the dataset?
Baskar Group (ComPM Lab) at Iowa State University, Ames, IA.

(b) How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The corresponding author of the main paper will be the single point of contact.

(c) Is there an erratum?
**[No]**

(d) Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances')?
**[No]**

(e) If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?
**[NA]**

(f) Will older versions of the dataset continue to be supported/hosted/maintained?
**[No]** . Since this is a physics dataset, we do not anticipate any changes to the content of our dataset.

(g) If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
**[No]**

(h) Any other Comments?
**[No]**

## 3 URLs for Dataset Access

- **Project Website:** https://baskargroup.bitbucket.io
- **Hosting Dataset:** https://figshare.com/s/15e9d23790d0a14e8f71
- **Code:** https://github.com/baskargroup/flowbench-tools

## 4 Author Responsibility Statement

As the authors of this submission, we affirm that we bear all responsibility in case of any rights violations or ethical issues associated with this work. We confirm that the submitted work is original, and if it includes third-party content (e.g., text, data, software), it is used with proper permissions and attributions. The authors assume full responsibility in case of violation of any rights.

### 4.1 Data License Confirmation

We confirm that all data used in this research has been generated by us. Furthermore, we confirm that the use of data in this research adheres to all applicable laws and ethical guidelines. In addition, we make our data available for general use under the terms of the CC BY-NC 4.0 license

## 5 Hosting, Licensing and Maintenance Plan

The authors are hosting all the data at the data release portal, with the URL available in the main paper. The data release portal is hosted by Iowa State University, Ames, IA, under an institutional license from Figshare. All the data made available will continue to be maintained by the authors at the aforesaid portal under the DOI - 10.25380/iastate.25939561

## References


[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

[2] Kumar Saurabh, Masado Ishii, Milinda Fernando, Boshun Gao, Kendrick Tan, Ming-Chen Hsu, Adarsh Krishnamurthy, Hari Sundar, and Baskar Ganapathysubramanian. Scalable adaptive pde solvers in arbitrary domains. In *Proceedings of the International Conference for high performance computing, networking, storage and analysis*, pages 1–15, 2021.

[3] Utkarsh Ayachit. *The paraview guide: a parallel visualization application*. Kitware, Inc., 2015.