FAERE

French Association
of Environmental and Resource Economists

Working papers

Prediction is difficult, even when it's about the past: a hindcast experiment using Res-IRF, an integrated energy-economy model

David Glotin - Cyril Bourgeois -
Louis-Gaëtan Giraudet - Philippe Quirion

WP 2019.03

Suggested citation:

D. Glotin, C. Bourgeois, LG. Giraudet, P. Quirion (2019). Prediction is difficult, even when it's about the past: a hindcast experiment using Res-IRF, an integrated energy-economy model.
FAERE Working Paper, 2019.03.

ISSN number: **2274-5556**

www.faere.fr

Prediction is difficult, even when it's about the past: a hindcast experiment using Res-IRF, an integrated energy-economy model

David Glotin (CIRED), Cyril Bourgeois (CIRED), Louis-Gaëtan Giraudet (CIRED, Ecole des Ponts Paris Tech), Philippe Quirion (CIRED, CNRS)

January 22, 2019

Keywords: retrospective simulation, backtesting, hindcast, model evaluation, model validation, buildings sector, residential sector.

Highlights

- We perform a retrospective simulation of an energy-economy model over 1984-2012.
- The model qualitatively replicates most observed trends but some observed evolutions are under-estimated: energy consumption per m² and the switch from fuel-oil to natural gas.
- We discuss possible explanations for the discrepancies; some of them are unlikely to be problematic for long-term projections.
- We conclude that hindcast experiments are useful to assess model performance and that understanding the causes of discrepancies with observations is essential to improve models performance.

Abstract

Model-based projections of energy demand are hardly ever confronted with observations. This shortfall threatens the credibility policy-makers might attach to integrated energy-economy models. One reason for it is the lack of historical data against which to calibrate models, a prerequisite for attempting to replicate past trends. In this paper, we (i) assemble piecemeal historical data to reconstruct the energy performance of the residential building stock of 1984 in France; (ii) calibrate Res-IRF, a bottom-up model of residential energy demand in France, against these data and run it to 2012. In a preliminary simulation that only considers the data that were known at the beginning of the simulated period, we find that the model accurately predicts energy consumption per m² aggregated over all dwelling types, with a Mean Absolute Percentage Error below 1.5% and 85% of the variance explained. These figures reach 0.5% and 96% when we consider the best-fit of 1,920 scenarios covering the uncertainty surrounding the parameters of the initial year. Energy demand is unevenly well replicated across fuels, which reveals some limitations in the ability of the model to capture politically-driven policies such as the expansion of the natural-gas distribution network. The overall results however build confidence in the general accuracy of the Res-IRF model. We discuss the directions for data collection which would ease comparison between simulations and observations in future hindcast experiments.

1. Introduction[1]

Prediction might be difficult when it's about the future,[2] but as we argue here, it can also be a daunting task when it comes to the past.

In most natural and engineering sciences, models are routinely compared to observations; cf. e.g. Legates and McCabe (1999) in hydrology, Brisson et al. (2002) in agronomy and Dudhia (1993) in climate science. For example, all climate models used for IPCC reports have been run using past forcing data (greenhouse gas concentrations, solar irradiance, volcanism…). International model intercomparison projects like CMIP[3] gather many modelling teams and provide them with a set of standard scenarios, tools and observational data, so that model outputs can be compared to one another, and to observations. Indeed comparing outputs of retrospective simulations to observations – a method referred to as "hindcasting", "backtesting", "predictive validation" or "cross-validation" – is essential to determine how well these models incorporate different parts of the climate system. This helps build confidence in model projections and identify model limitations that should be addressed in future research.

Such comparisons between retrospective simulations and observations are infrequent in economics, as highlighted by several researchers (Vallenzuela et al., 2007; Beckman et al., 2011; Baldos and Hertel, 2013; Northcott, 2019), in particular in energy economics. Energy-economy models include parameters calibrated or econometrically estimated on past data, but their ability to replicate past evolutions is rarely assessed (Beckman et al., 2011).

This scarcity of hindcasting exercises may be primarily explained by the lack of observational data against which to compare simulations, as noted by Chaturvedi et al. (2013). Moreover, the exercise has its own limitations, since the ability of a model to replicate past observations does not proves it a relevant tool for long-term projections (Calvin et al., 2017; Oreskes, 1998). Conversely, any inconsistency between simulations and observations at one point in time may be due to factors unlikely to happen again in the future, a point which we address in section 5 below.

Despite these limitations, the scarcity of hindcasting exercises has raised severe criticism. For instance, Kehoe (2005) writes "it is the responsibility of modelers to demonstrate that their models are capable of predicting observed changes, at least ex post. If a modeling approach is not capable of reproducing what has happened, we should discard it." Koomey et al. (2003) make the point more bluntly: "One of the most striking things about forecasters is their lack of historical perspective. They rarely do retrospectives, even though looking back at past work can both illuminate the reasons for its success or failure, and improve the methodologies of current and future forecasts".

A few hindcasting exercises based on economic models (mostly applied to energy, trade policies or agriculture and land-use) and published in the last ten years (cf. section 2 below) have started to close this research gap. We contribute to this burgeoning field by hindcasting and evaluating Res-IRF,

[1] For their useful comments, we thank Améline Vallet and Vincent Viguié.
[2] Apocryphal quote generally attributed to the physicist Niels Bohr (1885-1962).
[3] https://pcmdi.llnl.gov/?projects/cmip/index.php

an integrated energy-economy model of the French residential sector (Giraudet et al., 2012).[4] As this task was confronted to the above-mentioned lack of past data to initiate the model (the repartition of the dwelling stock by energy efficiency class), we had to reconstruct historical data in a preliminary step.

Reducing energy consumption and CO_2 emissions in this sector is essential to tackle climate change since the share of the residential sector in global final energy consumption amounts to 22%[5] in 2013 and "global building energy consumption could increase by 50% to 2050 without assertive energy efficiency action" (IEA, 2016). The Res-IRF model is meant to incorporate the main relevant policies (taxes, subsidies and regulations) and the main drivers of energy consumption in this sector: construction, destruction and thermal renovation flows, fuel switch and occupants' behaviour (e.g. thermostat setting).

We find that the model qualitatively replicates the main observed trends but quantitatively under-estimates some of them. This is true in particular for energy consumption per m² (except for dwelling oil-heated) and for the switch from fuel-oil to natural gas.

We then perform a sensitivity analysis on 8 parameters and select, out of 1 920 scenarios, the parameter combination underlying the scenario that best fits observations. We discuss possible explanations for remaining discrepancies and whether they would affect long-term projections in any relevant way. Some causes are unlikely to appear in the future while others could, and those should be focused on in future research.

The rest of the article is structured as follows. Section 2 reviews related hindcasting exercises conducted in the building sector. Section 3 presents the model, the observational data and the metrics we use to assess the model performance. Results are presented in section 4, Section 5 provides a broader discussion and section 6 concludes.

2. Literature review

To our knowledge, only three hindcasting exercises analysing energy consumption in the buildings or residential sector have been published: Chaturvedi et al. (2013), Fujimori et al. (2016) and van Ruijven et al. (2010).

Chaturvedi et al. (2013) evaluate the global integrated assessment model GCAM, and focus on its buildings component in the USA, one of its 32 regions. The authors calibrate the model for 1990 and compare simulation results to historical estimates for 1995, 2000, 2005 and 2010. They compare in particular the evolution of residential and commercial floorspace, as well as the energy demand disaggregated by fuel, US state and energy service. The model accurately replicates the observed increase in residential floor space. It also accurately replicates the observed decrease in residential heating final energy per m² (which is also a central output variable in our model), though at a slower

[4] This exercise is part of a broader research programme, initiated with a global sensitivity analysis of Res-IRF (Branger et al., 2015), which consists in evaluating the "quality", or "fitness-for-purpose," of the model (Oreskes, 1998).
[5] Calculation based on IEA (2016).

than observed rate. This is true for the three modelled heating fuels (gas, fuel-oil and electricity). As we shall see, our findings are very similar, a pattern we try to explain in section 5.

Another point made by the authors is “that the creation of a historical evaluation dataset is one of the foremost challenges in the evaluation process.” As explained above, we faced the same challenge, and thus agree that it certainly contributes to explaining the above-mentioned scarcity of retrospective studies in economics.

Fujimori et al. (2016) analyse the computable general equilibrium (CGE) model AIM. They compare the simulated and observed energy consumption, at the world and regional levels, from 1981 to 2010. For Europe, as shown is the online Supporting Information to their article, the simulated building energy demand is close to observations for all fuels taken together and for electricity taken in isolation, but less so for gas and solid fuels. Moreover the simulations do not replicate the observed downward trend in liquid fuels.

Van Ruijven et al. (2010) quantify uncertainty in the calibration of TIMER 2.0, a system dynamics model that simulates developments in global energy supply and demand. The authors focus on the effect of residential energy use at the regional scale for the period 1970-2003. For Western Europe, the model replicates the stability of residential fuel consumption for the period considered, but fuel consumption is not disaggregated across energies. It also replicates the observed growth in residential electricity consumption but not the slowdown observed in the second half of the period.

Some hindcasting exercises have also been conducted for energy-economy models which do not focus on the building sector.

Guivarch et al. (2009) use the oil price spike of 2008 to assess the Imaclim-R recursive CGE model. They compare the simulated and observed evolution of macroeconomic variables for a major oil importer (India). The model is found to overestimate the recessionary impact of the oil shock. The authors identify three mechanisms originally not included in their model, the inclusion of which reduces the gap between simulated and observed outputs. These three mechanisms are an increase in capital inflows, subsidies to domestic oil consumers and the rise of India as an exporter of refined products. The authors consider that these mechanisms are “bound to remain short-term” so “it appears acceptable not to embark these mechanisms in [their] modeling architecture when analysing long-term and global evolutions.” We discuss this point further in section 5 below.

Beckman et al. (2011) analyse the CGE model GTAP-E. By comparing the variance of model-generated petroleum price distributions – driven by historical demand and supply shocks to the model – with observed price distributions, they conclude that energy demand in GTAP-E is far too price-elastic. After incorporating the latest econometric estimates of energy demand and supply elasticities, they find the model to perform better.

Van Ruijven et al. (2009) apply the same analysis as Van Ruijven et al. (2010) but with a focus on transportation. They conclude that different model calibrations based on the same data lead to contrasted future projections, with a range in outcomes about 44–79% around the best-fit option.

In closing, we briefly mention some hindcasting exercises performed on economic models with no particular focus on energy. A first set of studies assessing the ability of CGE models to replicate the impact of trade agreements has produced contrasted conclusions (Kehoe et al., 1995; Kehoe 2005). A

second set of papers focus on agriculture and land-use (Baldos and Hertel, 2013; Calvin et al., 2017; Lotze-Campen et al., 2008; Ronneberger et al., 2008; Souty et al. 2013; Snyder et al., 2017; Valenzuela et al., 2007). On this topic, the Agricultural Model Intercomparison and Improvement Project (AgMIP; Rosenzweig et al., 2013) has the potential to facilitate the development of hindcasting exercises by mutualising methods, tools and datasets, like CMIP for global climate models.

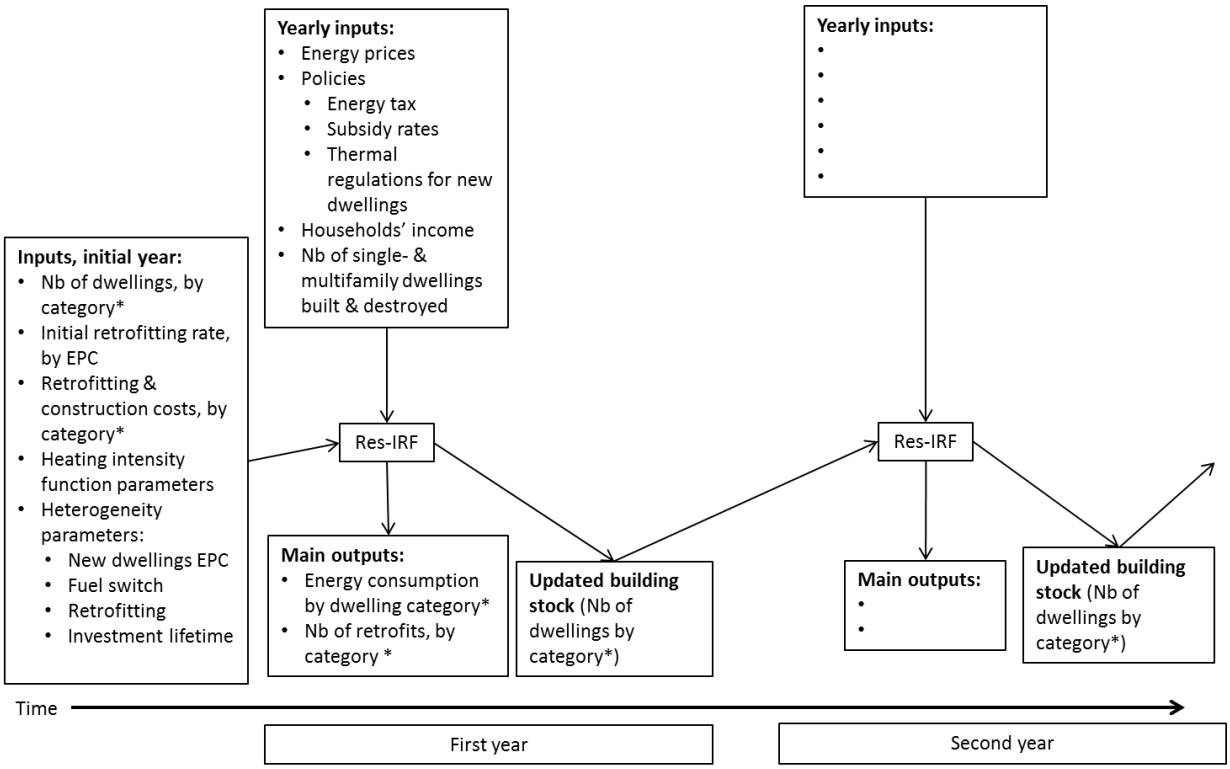
3. Material and methods

The Res-IRF[6] model

Res-IRF, described in Figure 1, is an integrated energy-economy model of energy use for heating in French dwellings developed at CIRED. Version 1 of the model is presented in Giraudet et al. (2011; 2012), version 2 in Branger et al. (2015) and version 3, the one used here, in Giraudet et al. (2018). It simulates the construction and destruction of new dwellings, thermal retrofitting of existing ones, the choice of the energy source in new dwellings, fuel switch in existing dwellings[7], and heating intensity (e.g. thermostat setting). The latter depends positively on household's income and dwelling energy efficiency, and negatively on the energy price. This implies that an energy tax will reduce heating intensity while a subsidy to energy efficiency will increase it (rebound effect).

[6] Res-IRF stands for the "Residential module of IMACLIM-R France". Res-IRF can indeed be linked to IMACLIM-R France, a CGE model developed at CIRED (Mathy et al., 2015). In this work, Res-IRF is run on its own.
[7] In the model, fuel switch can only occur when the dwelling is retrofitted. We come back to this point in section 5.

Figure 1. Simplified representation of the Res-IRF model



*A dwelling category is defined along 5 dimensions: EPC, household's income class, structural character, main heating energy, occupancy status.

Source: authors

Res-IRF is run at an annual time step and has a finer resolution (840 categories of dwellings) than that of those peer models that have also been subjected to hindcasting. In particular, Res-IRF distinguishes dwellings along five dimensions: tenancy status (owner-occupied, rented – private homeowner, rented – public homeowner), structural character (single-family vs. multi-family), main energy source (fuel-oil, natural gas, electricity and fuelwood[8]), household's income, and energy efficiency summarised by the energy performance certificate (EPC) class, from A to G, cf. Table 1. This facilitates the comparison of simulations and observations and the interpretation of observed discrepancies.

The model considers implicit technologies, i.e., it is not detailed if windows are double-glazed and walls have external insulation, but each change from a given energy performance class to a better one (entailed by thermal retrofitting) is associated to an investment costs in €/m². The cost transition matrix is presented in Giraudet et al. (2018). The share of dwellings being retrofitted and the ambition of this retrofitting (the EPC class reached) are represented by logit functions, which take into account the discounted cost of the available options, including the investment and energy cost

[8] We do not present results for dwellings heated mostly by fuelwood because the main part of fuelwood is used in dwellings using also another heating fuel, and determining the main energy source in these dwellings is problematic. This makes the comparison between model outputs and observations difficult. This problem does not significantly impact aggregate results since only half a million dwellings used wood as the main heating source in 2012 (CEREN, 2015).

(net of the various taxes and subsidies) and an intangible cost representing all the non-financial drivers of the investment (aesthetic or acoustic benefit, inconvenience due to the work…) that cannot be estimated. The intangible cost is calibrated so that the model reproduces a given retrofitting rate at the initial year (Giraudet et al., 2018).

Table 1. Energy consumption by Energy Performance Certificate class

EPC class	Energy consumption for heating (kWh/m²/yr.)
G	507
F	321
E	216
D	141
C	90
B	59
A	45

Conventional consumption, in primary energy. Source: authors' calculations, based on the Phébus survey.

The main time-varying exogenous inputs are energy prices, the number of households, households' income, and policies – thermal regulations, energy taxes and energy-efficiency subsidies (Table 2). In previous applications (forward-looking studies), Res-IRF model has been calibrated against year 2012 (using in particular the Phébus survey[9]) and run recursively in annual time steps to 2050 (Giraudet et al., 2018). For the calibration year, additional inputs are required, especially the state of the building stock, the costs of construction, of retrofitting and of fuel switch.

The different drivers of energy demand are determined endogenously (number of dwellings retrofitted, amount of energy saved thanks to these retrofits and occupants' behaviour), and several barriers to energy efficiency improvements are included (heterogeneity of consumer preferences, landlord-tenant dilemma, inertia of information diffusion, rebound effect). Positive externalities are also taken into account in the form of learning-by-doing (investment costs are reduced when the cumulated investments increase). Calibrated intangible costs mimic the imperfection of information and hidden costs that cannot be modelled explicitly.

Calibrating Res-IRF on the 1984-2012 period

In this work, the model was calibrated against 1984, a year chosen for data availability. Simulations were run until 2012 and outputs were compared to available observations over this period. Contrary to previous applications of Res-IRF, some demographic variables (number of households and share of multi-family vs. single-family dwellings) were set exogenously, based on observations[10] (Table 2).

[9] Enquête Performance de l'Habitat, Équipements, Besoins et Usages de l'énergie (Phébus). Ministère de la transition écologique et solidaire. http://www.statistiques.developpement-durable.gouv.fr/sources-methodes/enquete-nomenclature/1541/0/enquete-performance-lhabitat-equipements-besoins-usages.html

[10] In the first model runs (not reproduced here), the number of households and the share of multi-family vs. single-family dwellings were outputs of the model, determined by population and households' income.

The most demanding task was the assessment of the energy efficiency of the dwelling stock at the initial year (1984). Dwellings were classified in one of the EPC classes, according to the thresholds laid out in Table 1. It required the following information:

- The number of dwellings, split by heating fuel, income quintile, single vs. multi-family & tenancy status, was based on the National housing survey (Enquête nationale logement[11]).
- Realised final energy consumption per m², split by heating energy source and single vs. multifamily dwelling, was based on CEREN (2015).
- The "conventional energy consumption," i.e., the consumption for a standard utilisation of the heating system, was estimated based on the actual one, through the heating intensity function mentioned above.

Some exogenous variables were readily available: number of dwellings (split by heating energy source, income quintile & tenancy status), share of multi-family vs. single-family dwellings, share of owner-occupied vs. rented dwelling, households' income and energy prices. Table 2 provides the main data sources.

However compared to observations, the model overestimated the share of single-family dwellings, from around 2000 to 2012. This overestimation may be due to policies aimed at limiting urban sprawl, implemented at that time, including the law "solidarité et renouvellement urbain" voted in 2000 (Desrousseaux & Schmitt, 2018).

[11] https://www.insee.fr/fr/metadonnees/source/serie/s1004

Table 2. Main data sources

Data used as inputs to the model		
Variables	**Year(s)**	**Source**
Number of dwellings, split by heating energy source, income quintile & tenancy status	1984	National housing survey *(Enquête nationale logement)*
Energy prices	1984-2012	Pégase database[12]
Households' income	1984-2012	INSEE
Number of dwellings & share of single-family vs. multi-family	1984-2012	National housing survey *(Enquête nationale logement)*
Subsidy rates: reduced VAT, income tax credit for sustainable development (CIDD), Energy-Efficiency Certificates & zero-rate loan (éco-PTZ)	1984-2012	Various sources
Investment cost for new dwellings	1984-2012	INSEE, Cost of construction index
Data compared to model outputs		
Variables	**Year(s)**	**Source**
Repartition of dwellings by Energy Performance Certificate	2012	Phébus survey
Final energy consumption	1984-2012	CEREN
Number of dwellings, by energy source	1984-2012	CEREN

The least-known part concerned public policies (thermal regulations and subsidies) and especially the energy efficiency of the dwelling stock at the initial year (1984).

Thermal regulations for new residential buildings have been implemented since 1974 in France, and strengthened several times, with new regulations successively implemented for residential buildings in 1982, 1988, 2000, 2005 and 2012. Only the latter two specify explicit energy consumption limits. For the others, there is no direct correspondence between the requirements of the regulations and the EPC class. Besides, these regulations have been loosely enforced, at least up to the 1990s (Martin et al., 1998). The lack of ambition (or of enforcement) of these regulations is confirmed by the fact that in 2012, according to the above-mentioned Phébus survey, more than one million dwellings were classified as G, the worst EPC class. Based on these data and discussion with experts, we consider that dwellings built until 2004 are not influenced by thermal regulations while from 2005 onwards all new dwellings reach at least class E of EPC.

Reduced VAT has been in existence since 1999, income tax credit for sustainable development (CIDD) since 2005, Energy-Efficiency Certificates since 2006 and the zero-rate loan since 2009. These three policies are represented in the model as subsidies which reduce the investment cost of retrofitting.

[12] http://www.statistiques.developpement-durable.gouv.fr/donnees-ligne/r/pegase.html

Metrics to assess the model performance

In a first step, we present the results for the model calibrated without the post-1984 data[13] – thereafter referred to as the preliminary simulation. As highlighted by Calvin et al. (2017) among others, this is necessary to evaluate the performance of the model for prospective studies. Otherwise the model performance would be artificially boosted, compared to a situation where the model is actually used for forward projections.

To this end we then apply two widely-used (Bennet et al., 2013) evaluation metrics: the Mean Absolute Percentage Error (MAPE) and the coefficient of determination (R^2).

$$MAPE = \left(\frac{1}{n}\right) * \sum_{i=1}^{n} \frac{\left|X_{obs,i} - X_{model,i}\right|}{X_{obs,i}}$$

The MAPE is a frequently-used metric for hindcasting exercises (e.g. Calvin et al. 2017, Snyder et al., 2017) and represents the gap in percentage between simulations and observations, averaged over the period[14].

A limitation of the MAPE and similar indicators is that a low value may be due to the fact that observed variables vary little. In this case, a low MAPE only indicates that the model does not suffer from a systematic bias. Hence we also compute the coefficient of determination (R^2), following Lotze-Campen et al. (2008). R^2 elicits the share of the variance in the observed variable that is explained by the simulated one:

$$R^2 = \frac{\left(\sum_{i=1}^{n}\left(X_{obs,i} - \overline{X_{obs}}\right) * \left(X_{model,i} - \overline{X_{model}}\right)\right)^2}{\sum_{i=1}^{n}\left(X_{obs,i} - \overline{X_{obs}}\right)^2 * \sum_{i=1}^{n}\left(X_{model,i} - \overline{X_{model}}\right)^2}$$

with $\overline{X_{obs}}$ and $\overline{X_{model}}$ respectively the averages of observations and simulations and *n* the number of observations.

Both MAPE and R^2 are applied to final energy consumption per m^2, aggregated over all dwelling types, and over the period 1985-2012 for each scenario. We choose this output variable to compute the MAPE because it is the most policy-relevant among our model output variables.

In a second step, we perform a sensitivity analysis on those parameters that could not be precisely determined (table 3). All combinations are considered, which means that 1 920 ($2*2*2*3*2*5*4*2$) scenarios are run.

[13] Two caveats are in order. First, since res-IRF is a partial equilibrium model, some yearly variables are necessarily exogenous, in particular energy prices and household income. Second, as explained above, the number of households and the share of multi-family vs. single-family dwellings, which are endogenous when the model is run forward in time, are forced exogenously in this hindsight exercise so that the performance of the model with respect to more central outputs can be assessed.

[14] We prefer it to the RMSE (Root Mean Square Error), another frequently-used metric which is however known for putting too much weight on large errors (Bennett et al., 2013). Unlike the RMSE, the MAPE produces a non-smooth operator (it has a kink at zero) which is problematic in some optimisation contexts and explains why the RMSE is often preferred. This problem does not appear in our optimisation. Anyway, Snyder et al. (2017) show in a similar context that results are very close for the mean average error and for the RMSE.

Among the 1 920 simulations, we single out the "best fit" which minimizes the MAPE. The simulation which minimizes the other metric (the R^2) is very close so it is not presented. Parameters values are presented in Table 3 for the preliminary and the best-fit simulation.

Sensitivity analysis

In the best-fit simulation, compared to the preliminary one, the following input variables are changed. Investment costs become more expensive (for construction of new dwellings but also for retrofitting and fuel switch in existing dwellings); the heterogeneity parameter for new dwellings becomes higher, i.e. the EPC for new dwellings becomes more price-sensitive; heterogeneity parameters for both fuel switch and retrofitting become lower, i.e., the choice of an EPC and of a fuel becomes less price-sensitive; the retrofitting rate at the initial year (1984) becomes slightly higher; the Repartition of retrofitting by EPC includes less inefficient dwellings (EPC F & G) and more relatively efficient dwellings (EPC D & E); finally, to calculate the impact of retrofitting and fuel switch investments on energy consumption, shorter lifetimes are assumed, in other words investors do not take into account the increased value of the dwelling when they stop occupying them (i.e., after 7 years for owner-occupied dwellings and after 1 year for dwellings rented by a private homeowner).

Table 3. Summary of the sensitivity analysis

Input variable	Number of values	Value for the preliminary simulation	Value for the best-fit simulation
Investment cost for new dwellings	2	2012 value, inflation-adjusted	More expensive
Investment cost for existing dwellings (retrofitting & fuel switch)	2	2012 value, inflation-adjusted	More expensive
Heterogeneity parameter for new dwellings	2	8	15 (more price-sensitive)
Heterogeneity parameter for fuel switch	3	8	2 (less price-sensitive)
Heterogeneity parameter for retrofitting	2	8	2 (less price-sensitive)
Retrofitting rate at the initial year	5	3%	3.5%
Repartition of retrofitting by EPC	4	G: 40%; F: 35%; E: 20%; D: 5%; C,B,A: 0%	G:35%; F: 30%; E: 25%; D: 10%; C,B,A: 0%
lifetimes for retrofitting and fuel switch investments	2	With green values	Without green value

Source: authors

4. Results

Energy consumption per m²

Table 4, Figure 2 and figure 3 present the realised final energy consumption per m², the variable used to select the "best-fit" scenario. Results at the aggregated level are already satisfactory for the preliminary simulation since the Mean Absolute Percentage Error (MAPE) is lower than 1.5% and the model explains 85% of the variance (Table 4). Graphically, the decreasing trend is consistent with

observations (Figure 2). This trend may be explained by an increase in energy prices and the implementation of energy efficiency policies after 2000.

However, observed consumption decreases from around 150 kWh/m² in 1990 to around 100 in 2012 while in the preliminary simulation it starts from the same level but reaches around 120 kWh/m² in 2012. Results are much closer to observations with the best-fit simulation, which can be seen in Figure 2 (black curve) and by our model performance metrics: MAPE is divided by three, as well as the part of the variance not explained by the model (R^2 moves from 85% to 96%).

Table 4. Model performance metrics for the energy consumption per m²

	Mean Absolute Percentage Error (MAPE)		R^2	
	Preliminary simulation	**Best-fit simulation**	**Preliminary simulation**	**Best-fit simulation**
Aggregated	**1.4%**	**0.5%**	**85%**	**96%**
MF electricity	34.7%	27.3%	21%	95%
MF fuel-oil	2.7%	3.0%	56%	60%
MF natural gas	6.4%	3.2%	17%	60%
SF electricity	9.6%	3.7%	92%	86%
SF fuel-oil	2.7%	6.7%	70%	84%
SF natural gas	7.9%	2.0%	53%	96%

Source: authors

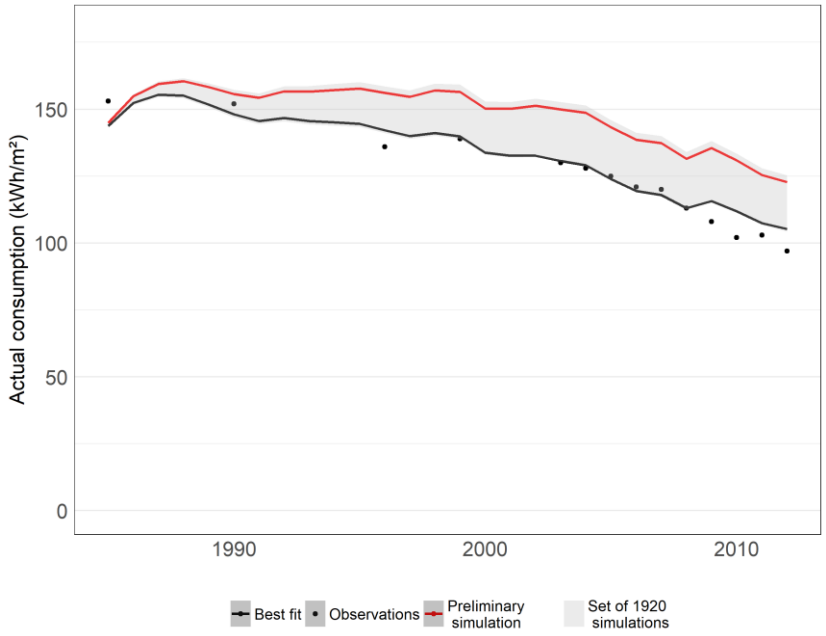


Figure 2. Realised final energy consumption for heating (kWh/m²) aggregated over all dwelling categories. Black dots: observations; grey zone: envelope of the simulations; red line: preliminary simulation; black line: best-fit simulation. Observations are weather-corrected (CEREN, 2015).

However disaggregated results are more contrasted. For each of the six dwelling categories, energy consumption decreases faster in the best-fit simulation than in the preliminary one. In most cases, the former is closer to observations but oil-heated dwellings are an exception on this point: the MAPE increases for these dwelling types, although the R^2 is slightly better. Moreover, all simulations largely underestimate electricity consumption in multi-family dwellings, especially at the end of the period; this bias is reflected in the high value of the MAPE for the category, in both the preliminary and the best-fit simulation.

The observed consumption of fuel-oil fluctuates more than that of other fuels, and not in line with our simulations. Consequently the MAPE is relatively high (6.7% in the best-fit simulation) and the R^2 relatively low (84%) for this dwelling category.

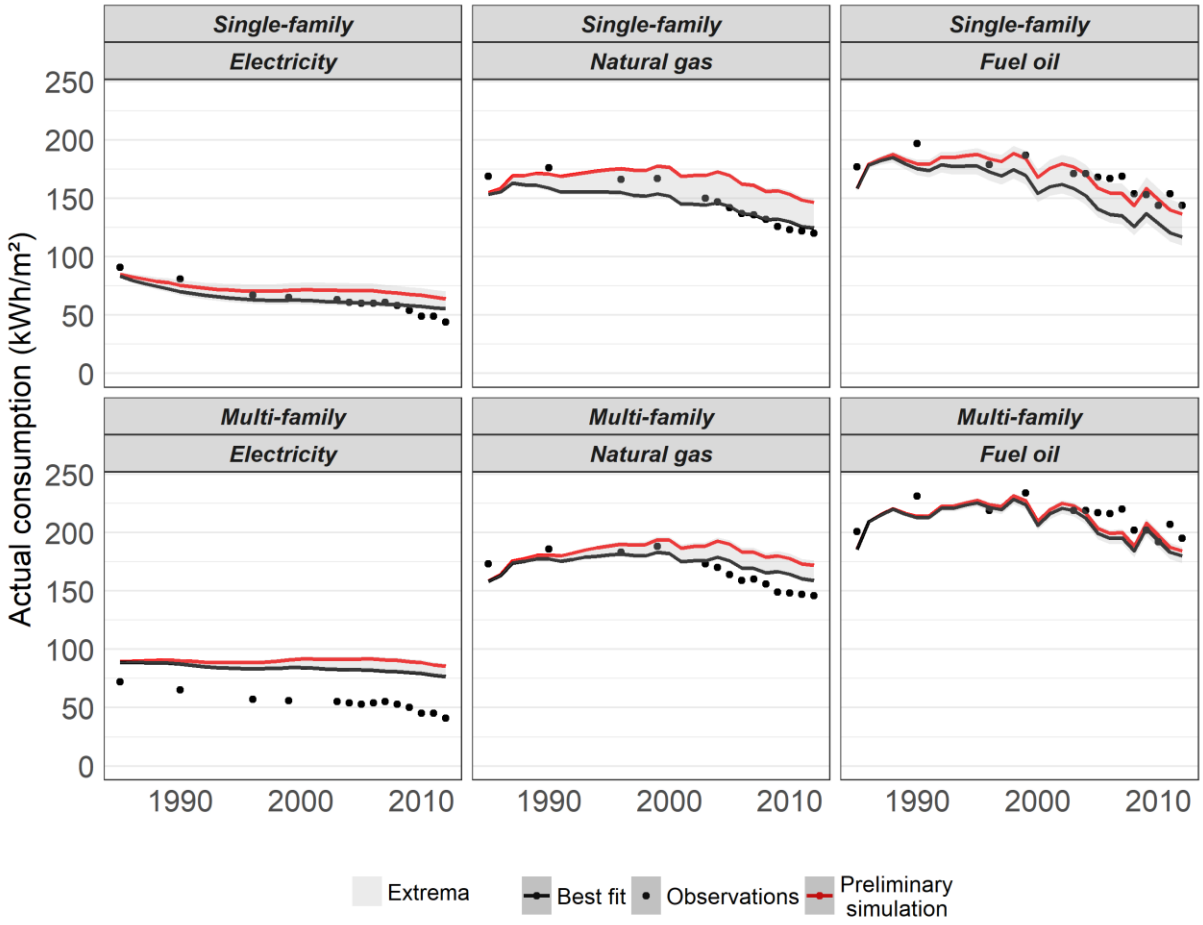


Figure 3. Realised final energy consumption for heating (kWh/m²), by dwelling category. Black dots: observations; grey zone: envelope of the simulations; red line: preliminary simulation; black line: best-fit simulation. Observations are weather-corrected (CEREN, 2015).

Stock evolution per type of dwelling and heating energy source

For each dwelling category, the model simulates the choice of the energy source, both for new and retrofitted buildings. Throughout the period, there has been a massive increase in the number of dwellings heated by natural gas and electricity and a drop in fuel-oil heating, especially in multi-family dwellings. Figure 3 shows that the model qualitatively replicates the observed trends. Quantitatively, it accurately replicates the rise in electric heating. For other fuels, simulations are accurate in trend but not in magnitude; specifically, the increase in natural gas and the decrease in fuel oil are largely underestimated. The best-fit simulation is not always closer to observations than the preliminary one, which is not illogical since this simulation has been selected for its capacity to fit the evolution of a different variable.

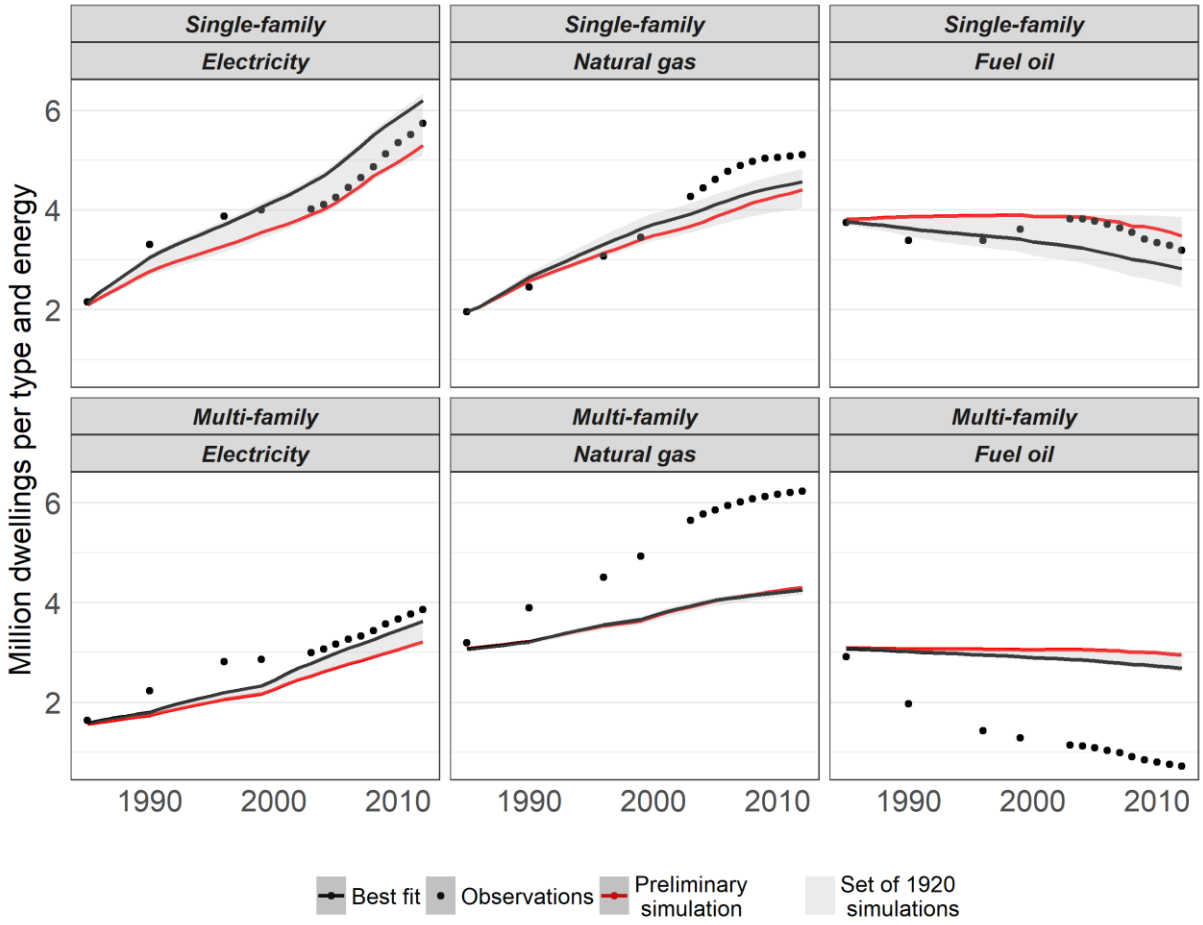


Figure 3. Evolution of the stock per type of dwelling and heating energy source

Energy performance of the dwelling stock

The only available data against which to compare our simulations come from the Phébus survey and concern a single year, 2012. However, the work realised to build the dwelling stock of 1984

generates insights into how the energy performance of dwellings has evolved between these two dates.

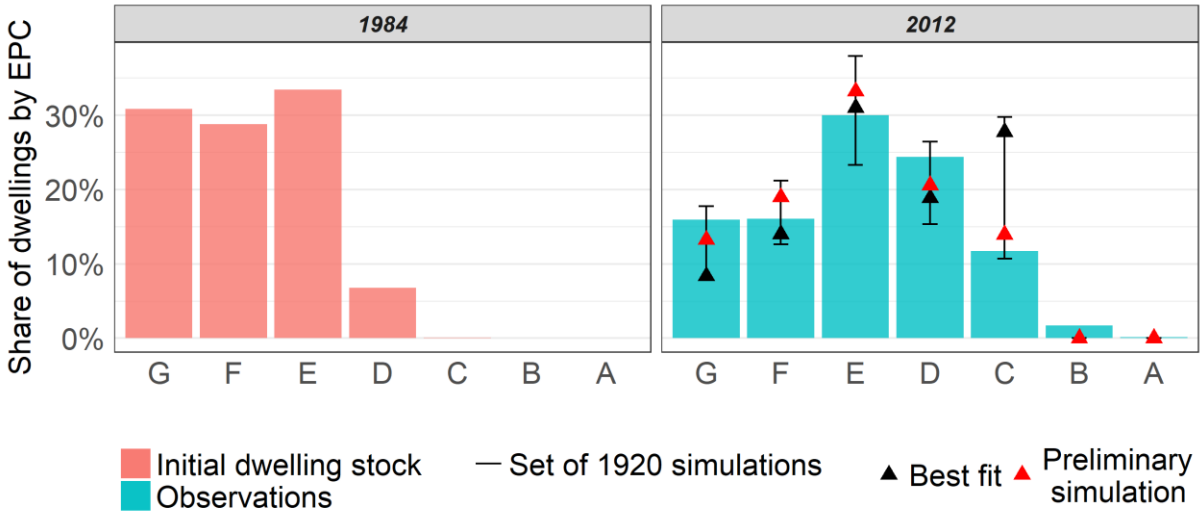


Figure 4. Distribution of dwellings per Energy Performance Class (EPC)

The left panel of Figure 4 represents the dwelling stock split by EPC of 1984, from the reconstitution explained in section 3. As we can see, there is almost no dwelling with an EPC A, B or C and only 7% of dwellings belong to class D.

On the right panel, observations from the Phébus survey are compared to simulations for the final year, 2012. In observations as well as in simulations, the dwelling stock is much more efficient in 2012 than in 1984, with especially a sharp decrease in class G dwellings. Simulations are also consistent with observations in that there is almost no dwelling belonging to classes A and B.

However, the simulated dwelling stock is more efficient than the observed one, with too many dwellings in class C and too little in class G. Moreover, for EPC classes C and G, the discrepancy is larger with the best-fit than with the preliminary simulation. In the best-fit simulation, the model parameters are set at values that permit particularly deep retrofits, thereby getting specific energy consumption closer to observations; this however comes at the expense of increasing the gap between the observed and simulated dwelling stock, in particular for the C and G EPC classes.

Number of retrofitted dwellings

Unfortunately we cannot compare the simulated number of retrofitted dwelling to observations for every year, nor for every dwelling category. The reason is that while estimations of the number of dwellings retrofitted have existed for several decades (Nauleau, 2014), until 2018 these surveys did not estimate the energy savings entailed by these retrofits. The newly released survey TREMI (Ademe, 2018) provide such data but only for single-family dwellings and only aggregated over the three years 2014-2016. Hence, we focus on single-family dwellings and on year 2012, the closest to the period covered by TREMI (Table 4).

Table 4. Thermal retrofitting of single-family dwellings: observations vs. simulations

	Observations (average 2014-2016)	Res-IRF, preliminary simulation (year 2012)	Res-IRF, best-fit simulation (year 2012)
Number of retrofitted single-family dwellings	433	301	271
Of which energy efficiency gain = 1 EPC class	347	175	153
Of which energy efficiency gain >= 2 EPC classes	87	125	117
Share of owner-occupied dwellings in retrofitted single-family dwellings	90%	86%	84%

Source: authors' calculations, based on Ademe (2018) for observations.

The number of retrofitted dwellings is about one third too low in the model simulations, compared to observations, the discrepancy being higher in the best-fit simulation. Among these retrofitted dwellings, those gaining 2 or more EPC classes are more numerous in the model simulations than in the observations. Finally, the dominance of owner-occupied dwellings in retrofitted dwellings (90%, while these dwellings represent 79% of single-family dwellings), is well represented.

These being said, these comparisons cannot be taken at face value, because they concern different years, and because the gain in EPC classes in TREMI is the result of a calculation based on a simplified building thermal model (3CL), which is itself imperfect and fed by rather imprecise data.

5. Discussion

Our hindcasting exercise indicates that the Res-IRF model is able to qualitatively replicate the main trends over the 1984-2012 period: the decrease in the number of dwelling heated by fuel-oil and the corresponding increase in the share of electricity and natural gas, the improvement in the building stock efficiency and the corresponding decrease in energy consumption per m² for each of the six dwelling categories (single-family and multi-family, heated by electricity, gas and fuel-oil). Moreover the simulated share of owner-occupied dwellings in retrofitted single-family dwellings is close to observations (table 4). In the remainder of this section, we discuss what we think are the main significant discrepancies.

Main discrepancies regarding aggregate energy consumption

In the preliminary simulation, the decrease in energy consumption per m² is lower than in observations. We can think of four explanations:

i. The number of retrofitted dwellings simulated could be too low. Two arguments point in this direction: in the best-fit simulation, the initial number of yearly retrofits (which is used to

calibrate intangible costs – cf. section 3 above) is higher than in the preliminary simulation (Table 3), and the decrease in energy consumption per m² is much closer to observations (Figure 2). Moreover the number of simulated retrofits in 2012 is lower than the one observed over 2014-2016 in average[15] (Table 4). A likely explanation for the too low number of retrofitted dwellings is that in the model, dwellings built after 1984 cannot be retrofitted, while in reality some of them have certainly been[16].

ii. Modelled retrofits could be less ambitious than in reality. Available observations do not point in this direction: first, the number of ambitious retrofits (at least two EPC classes gained) is higher than in observations (Table 4); second, the 2012 dwelling stock is too efficient compared to observations (Figure 4). Hence we think that we can rule out this explanation.

iii. The relatively coarse definition of the EPC classes may lead one to ignore some energy efficiency gains that do not generate a change in EPC class. In particular, the EPC class G has no upper bound for energy consumption. Hence the 31% dwellings falling in class G in 1984 may have benefited from significant energy retrofits without moving to a more efficient class. Also, while some dwellings built after 1984 undoubtedly fall in class G, they are certainly more efficient than the average class-G dwelling, built before the first thermal regulation in 1974.

iv. The heating intensity could increase too much throughout the period, perhaps because of a too high rebound effect. The rebound effect in Res-IRF is around 30%, a figure compatible with empirical estimates (Giraudet et al., 2018), but the latter typically yield large confidence intervals so we cannot rule out this explanation.

Main discrepancies by heating energy sources

The model vastly underestimates the switch from fuel-oil to natural gas in multi-family dwellings (Figure 3). We can think of two explanations.

i. One owes to the structure of the model, which allows fuel-switch only for dwellings which move to a better energy efficiency class, while in reality some dwelling have switched from fuel-oil to gas without improvement in the building envelope, and often without moving to a better energy class (even though new gas boilers generally had a better energy efficiency than the fuel boilers they replaced).

ii. Another explanation is that the switch was in practice enabled by the expansion of the natural gas network during the period simulated, especially in cities, where multi-family dwellings are concentrated. Absent geo-localised data of natural gas network and fuel-switch, we cannot quantify this effect.

The other main discrepancy concerns multi-family dwellings: those heated by fuel-oil and natural gas suffer from a relatively low R^2 (60% for the best-fit simulation) while those heated by electricity suffer from a relatively high MAPE (27% for the best-fit simulation) while observations indicate a much lower consumption than simulations, especially at the end of the simulation period (Figure 3). The latter discrepancy is particularly surprising and could be due to problems in observational data,

[15] Surprisingly, this number is even lower in the best-fit simulation; our interpretation is that in the best-fit simulation, more dwellings are retrofitted in the first years, so fewer opportunities for economically interesting retrofits are available at the end of the period.

[16] The TREMI survey indicates that 4% of the single-family dwellings retrofitted over 2014-2016 had been built after 2000 and 19% over 1975-2001 (Ademe, 2018). Unfortunately more precise data are not available.

since average energy consumption lower than 50 kWh of final energy per m² is surprisingly low even for dwellings heated by electricity, which are known to consume less than other dwellings, partly because energy is more expansive, partly because thermal regulations are more stringent.

Finally, as mentioned above, our simulations do not reproduce observed fluctuations in energy consumption of dwellings heated by fuel-oil. These fluctuations follow a pattern which seems difficult to explain; in particular it does not mirror fuel-oil price. Here again, this can be due to the imperfection of observational data, all the more likely that the sample of dwellings using fuel-oil for heating in CEREN data is smaller than for gas and electricity.

Implications of the main discrepancies for the model assessment

These differences between observations and simulations raise questions, which, we think, are quite general: to what extent do the discrepancies disqualify the use of the model for forward-looking studies? Do they point to prioritise some particular model developments? The answers clearly depend on the explanation for the observed discrepancies.

Take the first one, the too low decrease in energy consumption per m². As explained above, a likely explanation is that in the model, dwellings built after 1984 cannot be retrofitted (explanation i above). This is not problematic for forward-looking studies since dwellings built after 2012 are much more energy-efficient than before thanks to a new thermal regulation, so they are unlikely to be retrofitted in the next decades. Vice-versa, if the problem is a too high rebound effect (explanation iv), it will still happen in the future and deserves further attention. Finally, splitting class G in several sub-classes (following explanation iii) would be less useful for forward-looking studies since these dwellings represent a much smaller share now than in 1984.

Concerning the insufficient switch from fuel-oil to gas, the first explanation given above (fuel-switch only allowed for effective energy efficiency upgrades) may bias the results also for long-term projections, and could be addressed by changing the way fuel switch is modelled. In contrast, the second explanation (expansion of the gas network) is unlikely to bias long-term projections because the natural gas distribution network will not develop significantly in the future, since it has reached almost every densely built area[17].

To sum up, identifying discrepancies between model and observation does not systematically imply the model should be discarded or even modified – provided that the cause is unlikely to materialise again in the future. Guivarch et al. (2009) make a similar point when they write that the mechanisms behind the discrepancies they identified are bound to remain short-term, so are of little relevance for the long-term prospective studies their model is designed for.

Implications of the main discrepancies for modellers and data providers

Unfortunately, simulations are generally presented by researchers at a rather aggregate level which prevents one from further comparing models outputs. Systematically presenting energy consumption split by fuel and (when relevant) dwelling type would facilitate model comparisons for future hindcasting exercises; so would presenting energy consumption per m² in addition to aggregate

[17] Almost every municipality over 10 000 inhabitants is connected to the gas network. Connecting smaller municipalities generates disproportionate costs. https://www.ecologique-solidaire.gouv.fr/infrastructures-et-logistique-gazieres#e3

consumption. This could be facilitated by a model intercomparison project devoted to energy consumption, in the spirit of CMIP in climate science or AgMIP in the agriculture-land use scientific community. The Energy Modeling Forum could form a natural forum for this kind of retrospective studies since it already has this role for prospective studies (e.g. Huntington, 2011, in the field of energy efficiency).

However we think that the main hurdle to hindcasting exercises in energy economics is the limited availability of observational data. In France, yearly residential energy consumption data are available to researchers only in an aggregated way, individual data being private property of CEREN. This prevents researchers from disentangling the effect of problems in data from that of problems in the model. Similarly, data on retrofitted dwellings are available only for single-family dwellings and in an aggregated way, although the situation is likely to improve in the near future thanks to new surveys launched by Ademe and to a more open data policy.

6. Conclusion

Our hindcasting exercises indicate that the Res-IRF model is able to replicate the main trends over the 1984-2012 period: the decrease in the number of dwelling heated by fuel-oil and the corresponding increase in the share of electricity and natural gas, the improvement in the building stock efficiency and the corresponding decrease in energy consumption per m² for each of the six dwelling categories (single-family and multi-family, heated by electricity, gas and fuel-oil). The metrics calculated for the decrease in energy consumption per m² indicate a good overall performance of the model.

They also shed light on some limitations of the model, in particular too slow evolutions in fuel switch and energy consumption per m². This provides directions for improvements of the model in future work. To this end, it is useful to identify the cause of discrepancies between observations and simulations, because some of the possible causes are likely to persist in the future (like the possibility to switch fuel without improving the building envelope) while others are not (like the development of the natural gas network).

Data availability makes such identification difficult, a point also highlighted by Chaturvedi et al. (2013). Observations of energy consumption which we compare to our model outputs are based on a survey whose results are available to researchers only in aggregate. This hinders identification of the reasons for the observed discrepancies.

However we are convinced that these difficulties should not prevent economists to engage into hindcasting. To quote Oreskes et al. [42] “In areas where public policy and public safety are at stake, the burden is on the modeller to demonstrate the degree of correspondence between the model and the material world it seeks to represent and to delineate the limits of that correspondence.”

References

Ademe, 2018. *Enquête TREMI - Travaux de Rénovation Energétique des Maisons Individuelles. Campagne 2017*. ISBN 979-10-297-1023-0.

Baldos, U. L. C. and Hertel, T. W. (2013). Looking back to move forward on model validation: insights from a global model of agricultural land use, *Environ. Res. Lett.*, 8, 034024, https://doi.org/10.1088/1748-9326/8/3/034024, 2013.

Beckman, J., Hertel, T., and Tyner, W. (2011). Validating energy-oriented CGE models, *Energy Economics*, 33, 799–806.

Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., ... & Pierce, S. A. (2013). Characterising performance of environmental models. *Environmental Modelling & Software*, *40*, 1-20.

Branger, F, L.-G. Giraudet, C. Guivarch, P. Quirion (2015). Sensitivity analysis of an energy-economy model of the residential building sector. *Environmental Modelling & Software*. 70: 45-54.

Brisson, N., Ruget, F., Gate, P., Lorgeou, J., Nicoullaud, B., Tayot, X., ... & Mary, B. (2002). STICS: a generic model for simulating crops and their water and nitrogen balances. II. Model validation for wheat and maize. *Agronomie*, *22*(1), 69-92.

Calvin, K., Wise, M., Kyle, P., Clarke, L., & Edmonds, J. (2017). A hindcast experiment using the GCAM 3.0 agriculture and land-use module. *Climate Change Economics*, *8*(01), 1750005.

Chaturvedi V, et al. (2013) Model evaluation and hindcasting: An experiment with an integrated assessment model. *Energy* 61:479-490.

CEREN (2015). *Data on French dwellings from 1982 to 2015: number of dwellings by type and energy*. CEREN, Paris.

Desrousseaux, M. and B. Schmitt (2018). Réduire l'impact de l'artificialisation des sols. *L'économie Politique*, 2018/2 N° 78, pp. 54-68.

Dudhia, J. (1993). A nonhydrostatic version of the Penn State–NCAR Mesoscale Model: Validation tests and simulation of an Atlantic cyclone and cold front. *Monthly Weather Review*, *121*(5), 1493-1513.

Fujimori, S., Dai, H., Masui, T., and Matsuoka, Y. (2016). Global energy model hindcasting, *Energy*, 114, 293–301.

Giraudet, L.-G., C. Guivarch and P. Quirion (2012). Exploring the potential for energy conservation in French households through hybrid modelling, *Energy Economics*, 34: 426-445.

Giraudet, L.-G., C. Guivarch and P. Quirion (2011). Comparing and combining energy saving policies. Will proposed residential sector policies meet French official targets?, *Energy Journal*, 32(SI1): 213-242.

Giraudet, L.-G., C. Bourgeois, P. Quirion and D. Glotin (2018). Long-term efficiency and distributional impacts of energy saving policies in the French residential sector, CIRED, Paris. https://hal.archives-ouvertes.fr/hal-01890642/

Guivarch C, Hallegatte S, & Crassous R (2009) The resilience of the Indian economy to rising oil prices as a validation test for a global energy–environment–economy CGE model. Energy Policy 37(11):4259-4266

Huntington, H.G. (2011). The Policy Implications of Energy-Efficiency Cost Curves. *The Energy Journal* 32, 7–21.

IEA, 2016. *Energy Technology Perspectives*. International Energy Agency, Paris

Kehoe, T.J., 2005. An Evaluation of the Performance of Applied General Equilibrium Models of the Impact of NAFTA. Frontiers in Applied General Equilibrium Modeling: Essays in Honor of Herbert Scarf, ed. TJ Kehoe, TN Srinivasan, and J. Whalley. Cambridge, UK: Cambridge University Press.

Kehoe, T.J., Polo, C., Sancho, F., 1995. An evaluation of the performance of an applied general equilibrium model of the Spanish economy. Economic Theory 6, 115–141.

Koomey, J., Craig, P., Gadgil, A., & Lorenzetti, D. (2003). Improving long-range energy modeling: A plea for historical retrospectives. *The Energy Journal*, 75-92.

Legates, D. R., & McCabe Jr, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water resources research*, *35*(1), 233-241.

Lotze-Campen, H., Müller, C., Bondeau, A., Rost, S., Popp, A., and Lucht, W. (2008). Global food demand, productivity growth, and the scarcity of land and water resources: a spatially explicit mathematical programming approach, Agr. Econ., 39, 325–338. 6977, 6979

Martin, Y., et al. (1998). *La Maîtrise de l'énergie : rapport de l'instance d'évaluation*. Paris : La Documentation Française.

Mathy, S., Fink, M., & Bibas, R. (2015). Rethinking the role of scenarios: Participatory scripting of low-carbon scenarios for France. *Energy Policy*, *77*, 176-190.

Nauleau, M. L. (2014). Free-riding on tax credits for home insulation in France: An econometric assessment using panel data. *Energy Economics*, *46*, 78-92.

Northcott, R. (2019). Prediction versus accommodation in economics. *Journal of Economic Methodology*. doi: 10.1080/1350178X.2018.1561080

Oreskes, N. (1998). Evaluation (not validation) of quantitative models. *Environmental Health Perspectives*, *106*(Suppl 6), 1453.

Ronneberger, K., Berrittella, M., Boselle, F., and Tol, R. S. (2008). *KLUM@GTAP: Spatially-Explicit, Biophysical Land Use in a Computable General Equilibrium Model*, GTAP Working Paper No. 50, Center for Global Trade Analysis, Department of Agricultural Economics, Purdue University, 15 available at: https://www.gtap.agecon.purdue.edu/resources/download/3681.pdf

Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P., ... & Asseng, S. (2013). The agricultural model intercomparison and improvement project (AgMIP): protocols and pilot studies. *Agricultural and Forest Meteorology*, *170*, 166-182.

Ruijven B van, van der Sluijs JP, van Vuuren DP, Janssen P, Heuberger PSC, de Vries B. (2009). Uncertainty from model calibration: applying a new method to transport energy demand modelling. *Environ Model Assess*;15(3):175e88

Ruijven B van, de Vries B, van Vuuren DP, van der Sluijs JP (2010). A global model for residential energy use: uncertainty in calibration to regional data. *Energy*;35(1):269e82.

Souty, F., Dorin, B., Brunelle, T., Dumas, P., & Ciais, P. (2013). Modelling economic and biophysical drivers of agricultural land-use change. Calibration and evaluation of the Nexus Land-Use model over 1961–2006. Geoscientific Model Development Discussions, 6(4), 6975-7046.

Snyder, A. C., Link, R. P., & Calvin, K. V. (2017). Evaluation of integrated assessment model hindcast experiments: a case study of the GCAM 3.0 land use module. *Geoscientific Model Development (Online)*, *10*(PNNL-SA-125087).

Valenzuela, E., Hertel, T.W., Keeney, R., Reimer, J. (2007). Assessing global computable general equilibrium model validity using agricultural price volatility. *American Journal of Agricultural Economics* 89 (2), 383–397.